MTTTP1 Tilastotieteen johdantokurssi
Luento 6.9.2018

1 JOHDANTO

Tilastotiede menetelmätiede, joka käsittelee

- tietojen hankinnan suunnittelua

  otantamenetelmät, koejärjestelyt, kyselylomakkeet

- tietojen keruuta
- tietojen esittämistä

  kuvailevaa tilastotiedettä

- tietojen analysointia

  johtopäätelmien tekoa analysointimenetelmien avulla

Ks. myös

http://www.uta.fi/sis/mtt/uudet/MTT-CBDA-Peltonen-orientoivat_2015.pdf

http://fi.wikipedia.org/wiki/Tilastotiede

Soveltajat käyttävät tilastotieteilijöiden kehittämiä menetelmiä tietoaineiston

- keruuseen
- kuvailuun
- analysointiin

Tilastotiedettä käytetään hyväksi aina, kun käsitellään empiiristä tietoaineistoa. Tietotekniikka ja matematiikka ovat ”apuvälineitä”.

Tilastollinen analyysi voidaan karkeasti jakaa

- kuvailevaan analyysiin

  kuvataan tietoaineistoa, graafiset esitykset, tunnusluvut, taulukot

- tilastolliseen päättelyyn

  johtopäätelmät aineiston (otoksen) perusteella, todennäköisyyslaskentaan perustuvien tilastollisten testien ja analysointimenetelmien avulla

MTTTP1

- aineiston hankintaa
- aineiston sisältämän tiedon esittäminen
- tilastollisen testauksen alkeita

2 TILASTOLLINEN TUTKIMUS JA SEN TYÖVAIHEET

Populaatio

tutkimusobjektien muodostama joukko, johon tilastollinen tutkimus kohdistuu

Tilastoyksikkö eli havaintoyksikkö

populaatio yksikkö

Esim. Henkilö, kunta, valtio, ruokakunta, kirja, auto, liikenneonnettomuus, www-sivu tilastoyksiköitä

Empiirinen tutkimus tehdään lähes aina käyttäen vain osaa populaatiosta, otosta. Otoksen perusteella tehdään päättelyt koko populaatiosta.

Tilastoyksikön ominaisuudet tilastollisia muuttujia

Esim. Henkilön ikä ja sukupuoli, kunnan asukasluku, valtion sijainti, auton väri muuttujia

Yleisesti merkitään x, y, z, ..., x_1, x_2, x_3, ...

Empiirinen havaintoaineisto (*data*) saadaan mittaamalla tilastoyksiköiden ominaisuuksia.

Tilastolliset analyysimenetelmät ovat välineitä havaintoaineiston tutkimiseksi sekä johtopäätelmien tekemiseksi populaatiosta aineiston perusteella.

Esim. 2.1. Opintojaksolle ilmoittautuminen

– tilastoyksikkö opiskelija

– muuttujia

tutkinto-ohjelma

sukupuoli

opintojen aloitusvuosi

Esim. 2.2. Opintojakson tenttiin osallistujat

– tilastoyksikkö opiskelija

– populaatio esim. kaikki opintojakson opiskelijat

– muuttujia esim. opiskelijan tutkinto-ohjelma, tenttipisteet

Esim. 2.3.

a) Populaationa Suomen kunnat

– tilastoyksikkö kunta

– muuttujia esim.

kunnan asukasluku, asuntojen keskikoko, kunnan sijainti (maakuntaliitto)

b) Populaationa (tai otoksena) Eduskunta 2015

– tilastoyksikkö kansanedustaja

– muuttujia edustajan ikä, puolue, äänimäärä, ammatti

Esim. 2.4. Tapahtuma tilastoyksikkönä

- synnytys
- liikenneonnettomuus
- työtapaturma
- jääkiekko-ottelu

Tilastollisen tutkimuksen työvaiheet

1 Suunnittelu

- tutkimuskohteen & aiheen valinta
  - tilastoyksikkö
  - muuttujat
- tutkimuksen suorittamisen suunnittelu
  - kyselylomake
  - otantamenetelmä
  - koejärjestely jne.

2 Aineiston hankkiminen ja tallennus analysointia varten

- suunnitellun havaintoaineiston hankinta
- tallennus ja muokkaus analysointia varten

3 Aineiston kuvailu

– kuvailevan tilastotieteen keinoin aineiston sisältämän tiedon esittely ja tutkiminen

4 Tilastolliset mallit ja testaukset

– populaatiosta tehtyjen väittämien testaukset aineiston (otoksen) perusteella

– todennäköisyysteoriaan perustuvien tilastollisten mallien sovittaminen havaintoaineistoon

5 Raportointi

– johtopäätelmien teko ja niiden esittäminen ja tulkinta

Ks. Harjoitustyön ohjeet
http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/htyop118.pdf

Avainkäsitteet:

- Populaatio
- Otos
- Tilastoyksikkö
- Muuttuja
- Havaintoaineisto
- Tilastollinen tutkimus

MTTTP1, luento 11.9.2018

KERTAUSTA

- Populaatio
  tutkimusobjektien muodostama joukko, johon tilastollinen tutkimus kohdistuu, koko N

- Populaation yksikkö
  tilastoyksikkö, havaintoyksikkö

- Otos
  populaation osajoukko, koko n

- Tilastoyksikön ominaisuudet
  tilastollisia muuttujia

- Empiirinen havaintoaineisto (data)
  saadaan mittaamalla tilastoyksiköiden ominaisuuksia

- Tilastolliset analyysimenetelmät
  välineitä havaintoaineiston tutkimiseksi ja
  johtopäätelmien tekemiseksi

- Tilastollinen analyysi
  kuvailevaa analyysia
  tilastollista päättelyä

- Tilastollisen tutkimuksen työvaiheet

  1 Suunnittelu

  – tutkimuskohteen & aiheen valinta

    tilastoyksikkö

    muuttujat

  – tutkimuksen suorittamisen suunnittelu

    kyselylomake

    otantamenetelmä

    koejärjestely jne.

  2 Aineiston hankkiminen ja tallennus analysointia varten

  – suunnitellun havaintoaineiston hankinta

  – tallennus ja muokkaus analysointia varten

3 Aineiston kuvailu

– kuvailevan tilastotieteen keinoin aineiston sisältämän tiedon esittely ja tutkiminen

4 Tilastolliset mallit ja testaukset

– populaatiosta tehtyjen väittämien testaukset aineiston (otoksen) perusteella

– todennäköisyysteoriaan perustuvien tilastollisten mallien sovittaminen havaintoaineistoon

5 Raportointi

– johtopäätelmien teko ja niiden esittäminen ja tulkinta

Ks. Harjoitustyön ohjeet
http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/htyop118.pdf

3 HAVAINTOAINEISTO JA HAVAINTOMATRIISI

Aineiston hankinta

- otantatutkimus, päättely populaatiosta satunnaisesti populaatiosta tehdyn otoksen (satunnaisotoksen) perusteella

- kokeellinen tutkimus, päättely populaatiosta saatujen tulosten perusteella

Esim. 3.1. Päättelytilanteita
http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=7

a) Puolueen kannatuksen arviointi, esim. https://yle.fi/uutiset/3-10387592

Muodostetaan luottamusväli todelliselle kannatukselle.

b) Halutaan arvioida suomalaisten naisten keskipituutta. Lasketaan otoksesta keskipituus ja arvioidaan virhettä, joka liittyy päättelyyn. Tässä voidaan muodostaa keskipituudelle luottamusväli.

Otantamenetelmät (tapoja satunnaisotoksen tekemiseen)

- yksinkertainen satunnaisotanta YSO
- systemaattinen otanta SO
- ositettu otanta OO
- ryväsotanta RY

http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=7

Sopiva aineisto voi olla olemassa, se voidaan saadaan myös yhdistelemällä eri lähteistä.

Analysoitavassa aineistossa

- *n* tilastoyksikköä, a_1, a_2, a_3, … , a_n
- *p* muuttujaa, x_1, x_2, x_3, … , x_p

Havaintomatriisi on *n* x *p* –taulukko, jossa muuttujien arvot jokaiselta tilastoyksiköltä muodossa:

	x_1	x_2 …	x_j …	x_p
a_1	x_{11}	x_{12} …	x_{1j} …	x_{1p}
a_2	x_{21}	x_{22} …	x_{2j} …	x_{2p}
..				
a_i	x_{i1}	x_{i2} …	x_{ij} …	x_{ip}
.				
a_n	x_{n1}	x_{n2} …	x_{nj} …	x_{np}

Havaintomatriisissa *n* riviä ja *p* saraketta, sarake muodostaa kyseisen muuttujan jakauman.

Esim. CTESTI-aineisto, mikroluokkien verkossa

Muuttujia

IKÄ	Oppilaan ikä 1.11.1974 täysinä vuosina
PITUUS	Oppilaan pituus 1 cm tarkkuudella
PAINO	Oppilaan paino 100 g tarkkuudella.
HENGTIL	Hengitystilavuus 100 cm^3 tarkkuudella
COOPER	Cooperin testin tulos
LEUANVET	Leuanvetojen lukumäärä
KOULU	Oppilaan koulu MYP=0, TKYL = 1
LKASTE	8-luokkaisen koulun I, III, V, VII, Vastaavuus peruskoulun 5, 7, 9 ja lukion 2.
ÄIDINK1, ÄIDINK2, MATEM, KIELI1, KIELI2, KIELI3, LIIKUNTA	Ko. aineitten koulunumerot
SOSRYHMÄ	Isän ammatin koodaus 1-9, Rauhala: 1960-luvun suomalaisten yhteiskunnan sosiaalinen kerrostuneisuus ammatin arvostuksen valossa
KIPPI	Koodaus pääsee kipin= 1, ei pääse =0
PUOLIV	Koodaus pääsee = 1, ei pääse = 0
HAVOPP	Oppilaan numero aineistossa

Tutkimusongelmia?

Esim. PULSSI-aineisto
http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=102

Tutkimusongelmia?

Esim. 3.5.

HOTDOG-aineisto
http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=101

Muuttuja $/oz ilmoittaa unssihinnan dollareina

1 unssi = 28,35 g = 0,02835 kg, 1 dollari = 0,77€

Kilohinta = ($/oz)x0,77/0,02835.

http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=11

Tutkimusongelmia?

Esim. 3.3.

Myytyjä kiinteistöjä
http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=10

Muuttujia P = myyntihinta tuhansina dollareina

S = koko tuhansina neliöjalkoina

Eurohinta = 0,77 x P x 1000

Neliöt = 0,0929 x S x 1000

1 square foot = 0,0929 m^2

Neliöhinta = Eurohinta/Neliöt

Tutkimusongelmia?

Esim. 3.6.

Tampereella 12 kuukauden aikana myytyjä kerrostaloasuntoja, otos 4.6.2012, aineisto Tre_myydyt_asunnot_2012.sav sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

Tutkimusongelmia?

4 MITTAAMINEN

- Mittaaminen

  menettely (sääntö), jolla tilastoyksikköön liitetään tiettyä ominaisuutta kuvaava luku, mittaluku.

- Mittausvirhettä

  mittari epätarkka

  häiriötekijät

- Mittarin reliabiliteetin alhainen

  toisistaan riippumattomat, samalle tilastoyksikölle tehdyt mittaukset antavat huomattavasti poikkeavia tuloksi

- Mittarin ei validi
  ei mittaa sitä ominaisuutta, mitä tarkoitus mitata (mittari huonosti laadittu)

- Suoraan mitattavissa ja tulkittavissa olevia muuttujia
  Esim. Henkilön pituus, paino, kunnan asukasluku, lasten lukumäärä perheessä

- Eivät suoraan mitattavissa olevia muuttujia, määrittely ei yksikäsitteistä
  Esim. Henkilön älykkyys, musikaalinen lahjakkuus, uskonnollisuus, asenne johonkin; www-sivun käytettävyys

Esim. Henkilön uskonnollisuutta voidaan mitata kirkossa käyntien määrällä, uskonnollisen kirjallisuuden lukemisella, ...(nk. indikaattorimuuttujien avulla).

Esim. Asenne-/mielipidemittauksissa asennetta/mielipidettä peilaavia väitteitä

Vastaaja valitsee esimerkiksi vaihtoehdoista

- □ täysin samaa mieltä
- □ jokseenkin samaa mieltä
- □ ei samaa eikä eri mieltä
- □ jokseenkin eri mieltä
- □ täysin eri mieltä

Muuttujia voidaan luokitella monella tavalla:

1) kategorisiin eli kvalitatiivisiin
   numeerisiin eli kvantitatiivisiin

2) mitta-asteikkojen perusteella

3) jatkuva
   ei-jatkuva

4) selitettävä
   selittäjä

1)

$\boxed{\text{Kvalitatiivinen (kategorinen) muuttuja}}$

jakaa tilastoyksiköt tarkasteltavan ominaisuuden suhteen luokkiin

Esim. Henkilön siviilisääty, opiskelijan tutkinto-ohjelma, kaupungin sijaintikunta, vaatteiden kokoluokitus

Kvalitatiiviset muuttujat voidaan koodata numeerisesti, MUTTA numeroarvoilla ei määrällistä tulkintaa; ovat vain luokkien nimiä tai kuvaavat luokkien "suuruusjärjestyksen".

$\boxed{\text{Kvantitatiivinen (numeerinen) muuttuja}}$

muuttujan arvo mitattaessa reaalinen, mitataan lukumäärää tai mittaus mittayksikköä käyttäen

Esim. Henkilön pituus, opiskelijan ikä, kaupungin asukasluku, vaatteen hinta

2)
Muuttujien mitta-asteikot

$\boxed{\text{Luokittelu- eli laatuero- eli nominaaliasteikko}}$

kvalitatiivinen muuttuja, jonka luokkia ei voida asettaa järjestykseen (esim. paremmuus, suuruus, kovuus)

Esim. Henkilön siviilisääty, opiskelijan koulutusohjelma, kaupungin sijainti

Järjestys- eli ordinaaliasteikko

kvalitatiivinen muuttuja, jonka luokat voidaan asettaa mielekkääseen järjestykseen mitattavan ominaisuuden suhteen

Esim. Asennekysymykset, vaatteiden kokoluokitus

Suhdeasteikko

numeerisen muuttuja, jonka arvo nolla vastaa tarkasteltavan ominaisuuden “häviämistä”, absoluuttista nollapistettä

Esim. Henkilön paino (kg) ja pituus (cm), henkilön 100 m juoksuaika (s), asunnon vuokra (€), urheilijan harjoitteluun käyttämä aika päivässä (min)

Intervalliasteikko

numeerisen muuttuja, jonka nollakohta ei suhdeasteikon tapaan määritelty

Esim. Huoneen lämpötila Celsius-asteina.

Absoluuttinen asteikko

suhdeasteikollinen, jossa mittaus kiinnitetyllä mittayksiköllä

Esim. Asunnon huoneiden lukumäärä, perheessä lasten lukumäärä

Esim. 4.2.

Liikuntamäärien mittaus

http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=14

Kvantitatiivisesti mitattuna:

Harrastan liikuntaa (väh. 30 min/kerta) keskimäärin ____ kertaa viikossa.
Keskimäärin kerralla liikun ____ min.

Kvalitatiivisesti mitattuna:

Harrastan liikuntaa (väh. 30 min/kerta) keskimäärin
____ alle 3 kertaa viikossa
____ 3–4 kertaa viikossa
____ enemmän kuin 4 kertaa viikossa.

Keskimäärin kerralla liikun
____ alle tunnin
____ 1–2 tuntia
____ yli 2 tuntia

Esim. CTESTI-aineiston muuttujien mitta-asteikot

IKÄ	Oppilaan ikä 1.11.1974 täysinä vuosina
PITUUS	Oppilaan pituus 1 cm tarkkuudella
PAINO	Oppilaan paino 100 g tarkkuudella.
HENGTIL	Hengitystilavuus 100 cm^3 tarkkuudella
COOPER	Cooperin testin tulos
LEUANVET	Leuanvetojen lukumäärä
KOULU	Oppilaan koulu MYP=0, TKYL = 1
LKASTE	8-luokkaisen koulun I, III, V, VII, Vastaavuus peruskoulun 5, 7, 9 ja lukion 2.
ÄIDINK1, ÄIDINK2, MATEM, KIELI1, KIELI2, KIELI3, LIIKUNTA	Ko. aineitten koulunumerot
SOSRYHMÄ	Isän ammatin koodaus 1-9, Rauhala: 1960-luvun suomalaisten yhteiskunnan sosiaalinen kerrostuneisuus ammatin arvostuksen valossa
KIPPI	Koodaus pääsee kipin= 1, ei pääse =0
PUOLIV	Koodaus pääsee = 1, ei pääse = 0
HAVOPP	Oppilaan numero aineistossa

Mitta-asteikko vaikuttaa tilastollisen menetelmän valintaan. Numeeristen muuttujien yhteydessä lähes samat menetelmät ja tunnusluvut käyvät kaikille kolmelle mitta-asteikolle.

Suhdeasteikolla muuttujan arvojen suhteilla on mielekäs tulkinta. Intervalliasteikolla voidaan vertailla arvojen eroja, mutta ei suhteita.

Avainkäsitteet:

- Havaintomatriisi
- Muuttujan jakauma
- Otantamenetelmät
- Mittaaminen
- Kvalitatiivinen muuttuja
- Kvantitatiivinen muuttuja
- Mitta-asteikot

MTTTP1, luento 13.9.2018

KERTAUSTA

- Havaintomatriisi

Tilastoyksiköt: $a_1, a_2, a_3, \ldots, a_n$
Muuttujat: $x_1, x_2, x_3, \ldots, x_p$
Havaintomatriisissa *n* riviä ja *p* saraketta

	x_1	x_2 ...	x_j ...	x_p
a_1	x_{11}	x_{12} ...	x_{1j} ...	x_{1p}
a_2	x_{21}	x_{22} ...	x_{2j} ...	x_{2p}
..				
a_i	x_{i1}	x_{i2} ...	x_{ij} ...	x_{ip}
.				
a_n	x_{n1}	x_{n2} ...	x_{nj} ...	x_{np}

- Mitta-asteikot

  Nominaaliasteikko
  Järjestysasteikko
  Intervalliasteikko
  Suhdeasteikko
  Absoluuttinen asteikko

  http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/mitta_asteikot_kuva.pdf

Esim. CTESTI-aineiston muuttujien mitta-asteikot

IKÄ	Oppilaan ikä 1.11.1974 täysinä vuosina
PITUUS	Oppilaan pituus 1 cm tarkkuudella
PAINO	Oppilaan paino 100 g tarkkuudella.
HENGTIL	Hengitystilavuus 100 cm^3 tarkkuudella
COOPER	Cooperin testin tulos
LEUANVET	Leuanvetojen lukumäärä
KOULU	Oppilaan koulu MYP=0, TKYL = 1
LKASTE	8-luokkaisen koulun I, III, V, VII, Vastaavuus peruskoulun 5, 7, 9 ja lukion 2.
ÄIDINK1, ÄIDINK2, MATEM, KIELI1, KIELI2, KIELI3, LIIKUNTA	Ko. aineitten koulunumerot
SOSRYHMÄ	Isän ammatin koodaus 1-9, Rauhala: 1960-luvun suomalaisten yhteiskunnan sosiaalinen kerrostuneisuus ammatin arvostuksen valossa
KIPPI	Koodaus pääsee kipin= 1, ei pääse =0
PUOLIV	Koodaus pääsee = 1, ei pääse = 0
HAVOPP	Oppilaan numero aineistossa

5 EMPIIRISET JAKAUMAT

5.1 Yksiulotteinen jakauma

Havaintomatriisin sarakkeilla on muuttujien x_1, x_2, x_3, ..., x_p jakaumat.

5.1.1 Frekvenssijakauma

Esim. Opintojaksolle ilmoittautuneet, $n = 442$, 10.9.

Opiskelijat tutkinto-ohjelmittain (%)

Matematiikka & Tilastotiede	10
Tietojenkäsittelytieteet	31
Kauppatieteet	24
Hallintotieteet	20
Muut	15

Tutkinto-ohjelmaopiskelijoista ensimmäisen vuoden opiskelijoita 17 %

Esim. 5.1.2. Yritykset toimialoittain

Toimiala	Frekvenssi
Elintarviketeollisuus	7
Ilmailu	18
IT	22
Lääketeollisuus	12
Paperinjalostus	11
Öljynjalostus	27
	97

Esim. 5.1.1. Lepopulssin jakauma, PULSSI-aineisto, liite 4

Pulssi -muuttujan luokat	Tilastoyksiköiden määrä luokassa
42–52	4
53–63	7
64–74	31
75–85	25
86–96	12
97–107	1

Esim. 5.1.5 ja 5.1.6 Lepopulssin mittaustarkkuus 1, pyöristetyt luokkarajat, todelliset luokkarajat, luokkakeskukset, luokan pituus 11, frekvenssit, summafrekvenssit

pyöristetyt luokkarajat	todelliset luokkarajat	Luokkak.	Frekv.	Summafrekv.
42–52	41,5–52,5	47	4	4
53–63	52,5–63,5	58	7	11
64–74	63,5–74,5	69	31	42
75–85	74,5–85,5	80	25	67
86–96	85,5–96,5	91	12	79
97–107	96,5–107,5	102	1	80

Frekvenssit ja summafrekvenssit voidaan esittää myös prosentteina.

Esim. 5.1.11. Miesopiskelijoiden pituus.

pyöristetyt luokkarajat	frekv.	summa-frekv.	luokka-keskus	todelliset luokkarajat
154-160	5	5	157	153,5-160,5
161-167	20	25	164	160,5-167,5
168-174	39	64	171	167,5-174,5
175-181	28	92	178	174,5-181,5
182-188	8	100	185	181,5-188,5

Mittaustarkkuus 1, luokan pituus 7

Ehdolliset frekvenssijakaumat, tarkastellaan frekvenssijakaumaa toisen muuttujan mukaan ryhmiteltynä.

Esim. 5.1.7. Lepopulssin jakauma miehillä ja naisilla

	Mies	Nainen
42–52	4 9,1 %	0 0,0 %
53–63	6 13,6 %	1 2,8 %
64–74	18 40,9 %	13 36,1 %
75–85	13 29,5 %	12 33,3 %
86–96	3 6,8 %	9 25,0 %
97–107	0 0 %	1 2,8 %
	44	36

Esim 5.1.4. Tampereen yliopistosta maisterin tutkinnosta 2016 valmistuneiden työelämään sijoittuminen, työtilanne koulutusaloittain vuosi valmistumisen jälkeen

https://www.uta.fi/opiskelunopas/tyoelama/valmistuneet-tyoelamassa

https://intra.uta.fi/portal/documents/159280/44060654/sijoittumisseuranta+2016.pdf/71ca38b5-90a6-4378-bbbf-1e697853e49a (s. 14)

Esim. Huoneiden lukumäärä alueittain, aineisto Tre_myydyt_asunnot_2012.sav sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

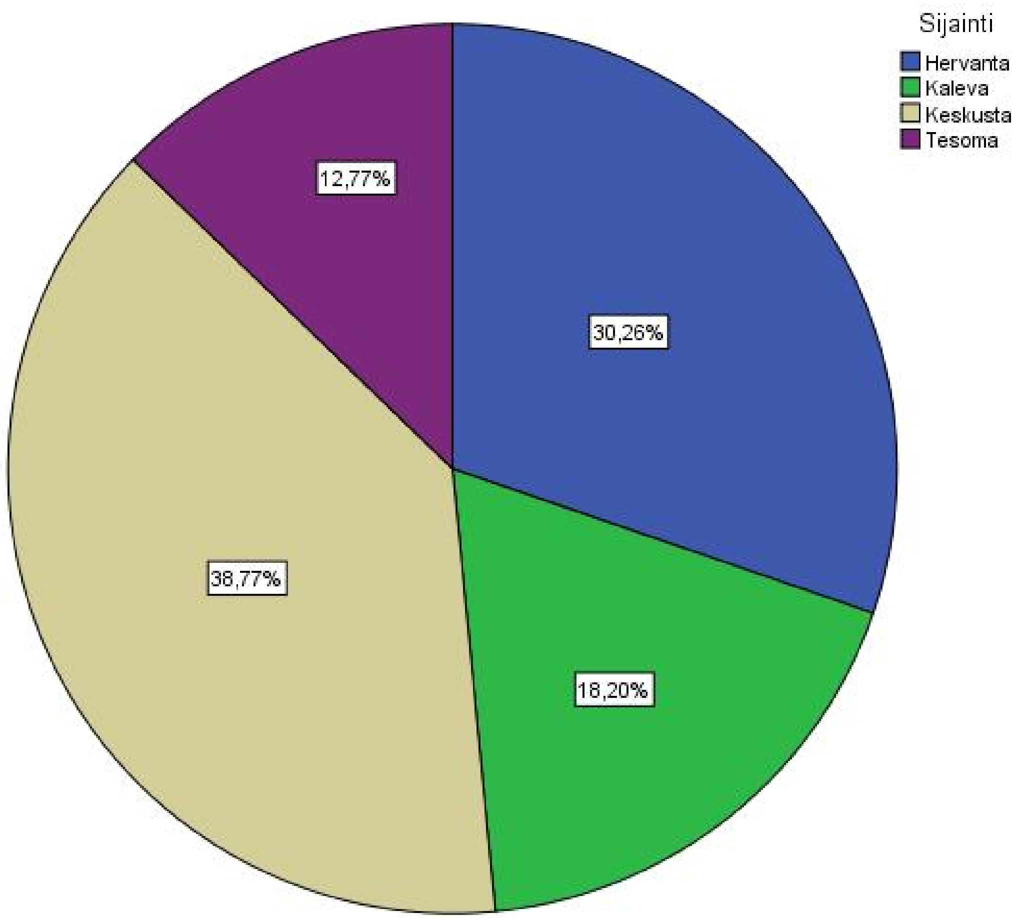
			Sijainti				Total
			Hervanta	Kaleva	Keskusta	Tesoma	
Huoneiden lukumäärä	1	Count	21	15	45	5	86
		% within Sijainti	16,4%	19,5%	27,4%	9,3%	20,3%
	2	Count	53	43	66	14	176
		% within Sijainti	41,4%	55,8%	40,2%	25,9%	41,6%
	3	Count	42	15	41	19	117
		% within Sijainti	32,8%	19,5%	25,0%	35,2%	27,7%
	4	Count	12	4	12	16	44
		% within Sijainti	9,4%	5,2%	7,3%	29,6%	10,4%
Total		Count	128	77	164	54	423
		% within Sijainti	100,0%	100,0%	100,0%	100,0%	100,0%

5.1.2 Frekvenssijakaumien graafiset esitykset

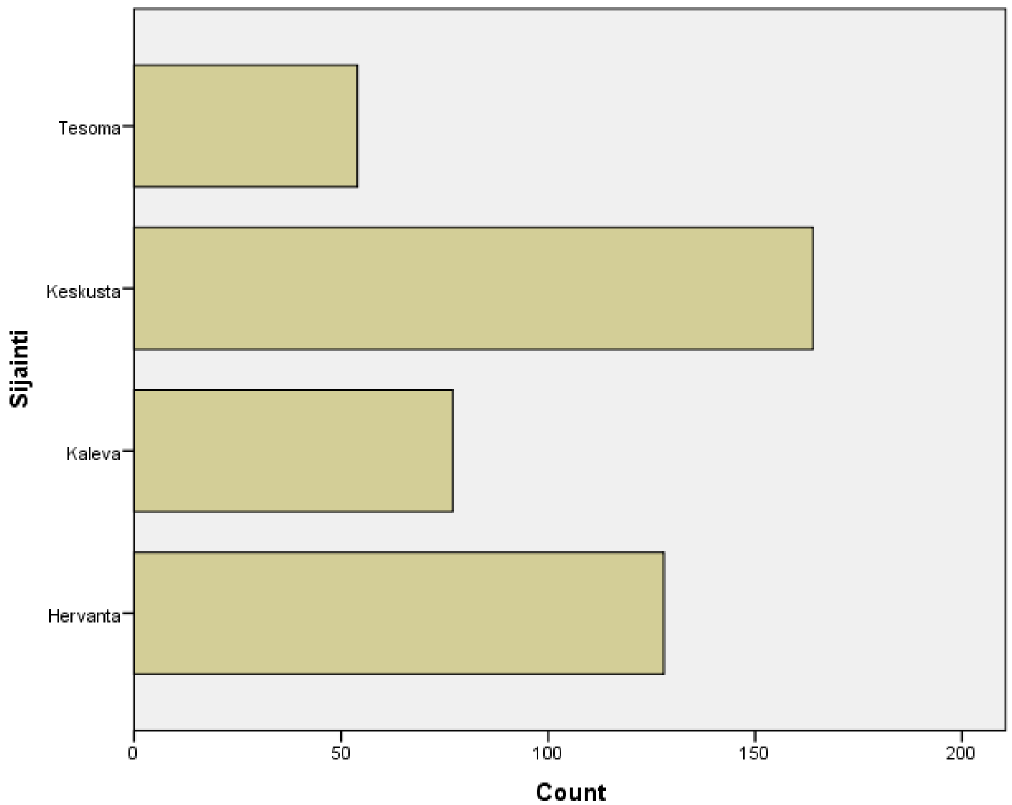
- Piirakkakuvio
- Pylväs-, vaakapylväs-, janadiagrammi
- Frekvenssihistogrammi

Esim. Aineisto Tre_myydyt_asunnot_2012.sav, muuttujien sijainti, huoneiden lukumäärä, neliöhinta graafiset esitykset ovat piirakkakuvio tai vaakapylväsdiagrammi, pylväsdiagrammi tai janadiagrammi, frekvenssihistogrammi. Neliöhintaa voidaan tarkastella myös sijainnin mukaan ehdollistettuna samoin huoneiden lukumäärää.

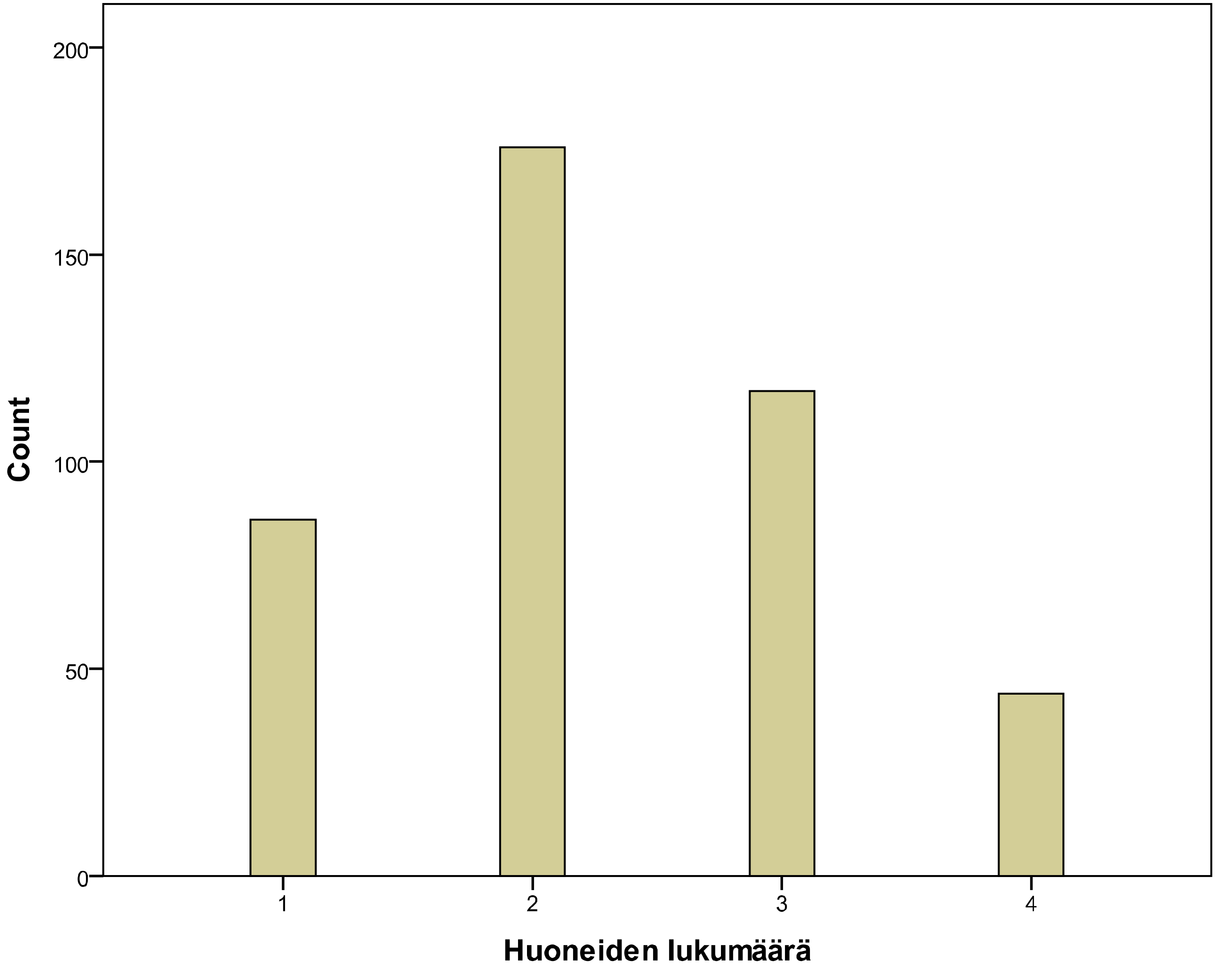
- Sijainnin jakauma, piirakkakuvio



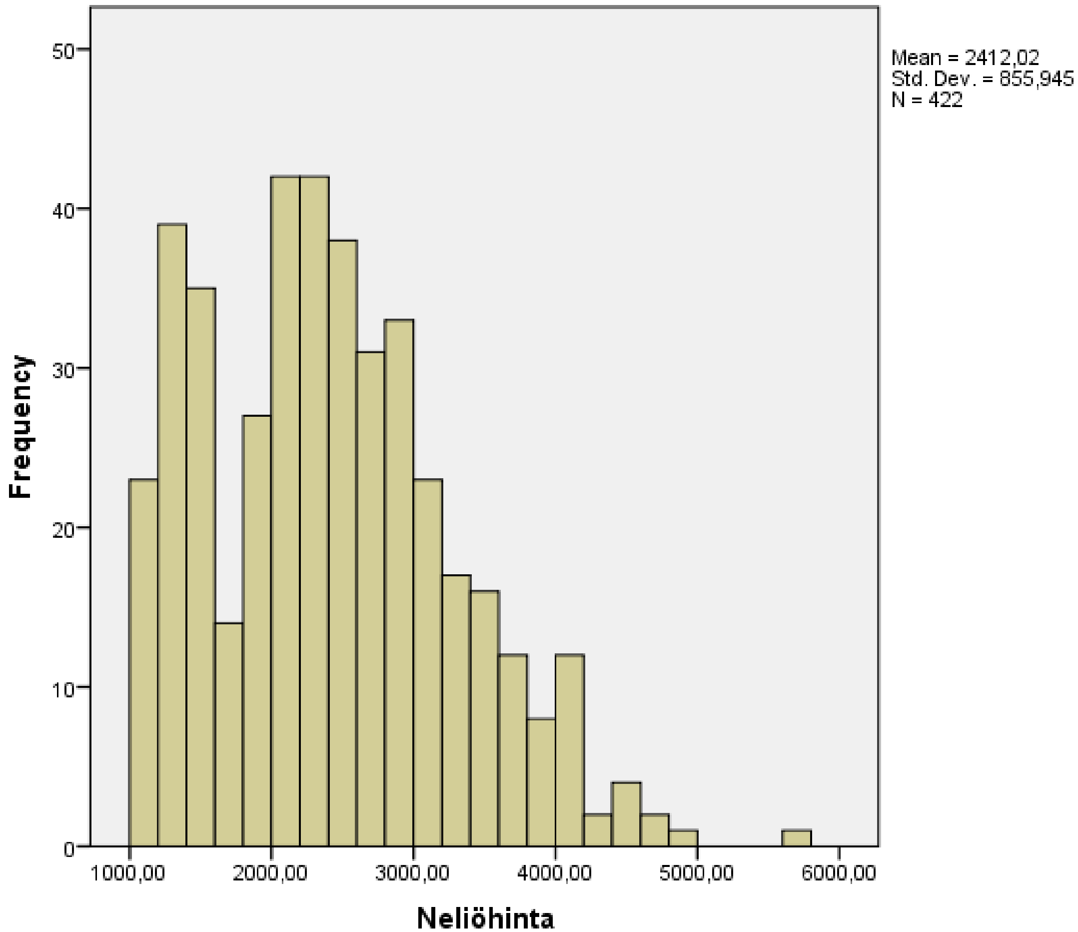
- Sijainnin jakauma, vaakapylväsdiagrammi



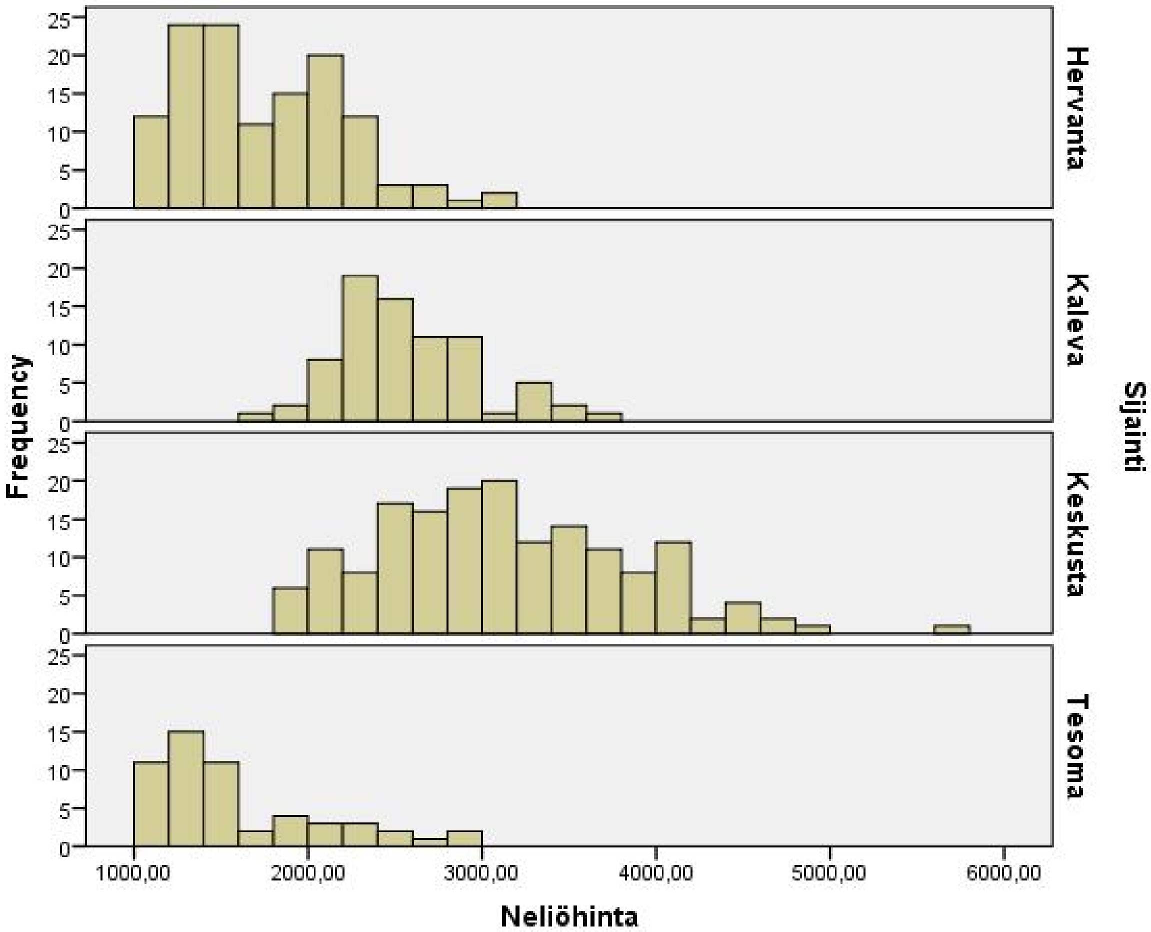
- Huoneiden lukumäärän jakauma, pylväsdiagrammi



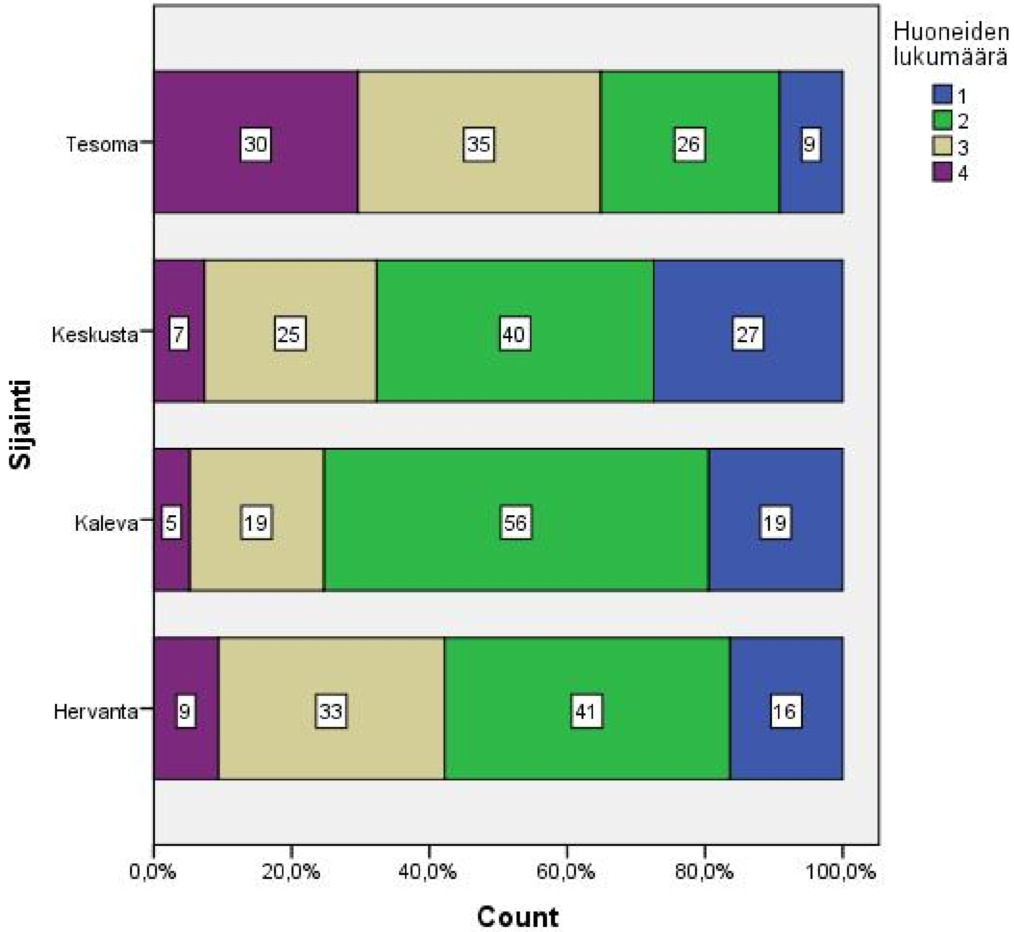
- Neliöhinnan jakauma, frekvenssihistogrammi



- Neliöhinnan ehdolliset histogrammit



- Huoneiden lukumäärä sijainnin mukaan, summapylväsdiagrammi



Grafiikan valinnasta esimerkkejä ks.
http://www.sis.uta.fi/tilasto/mtttp1/syksy2013/grafiikan_valinnasta.pdf

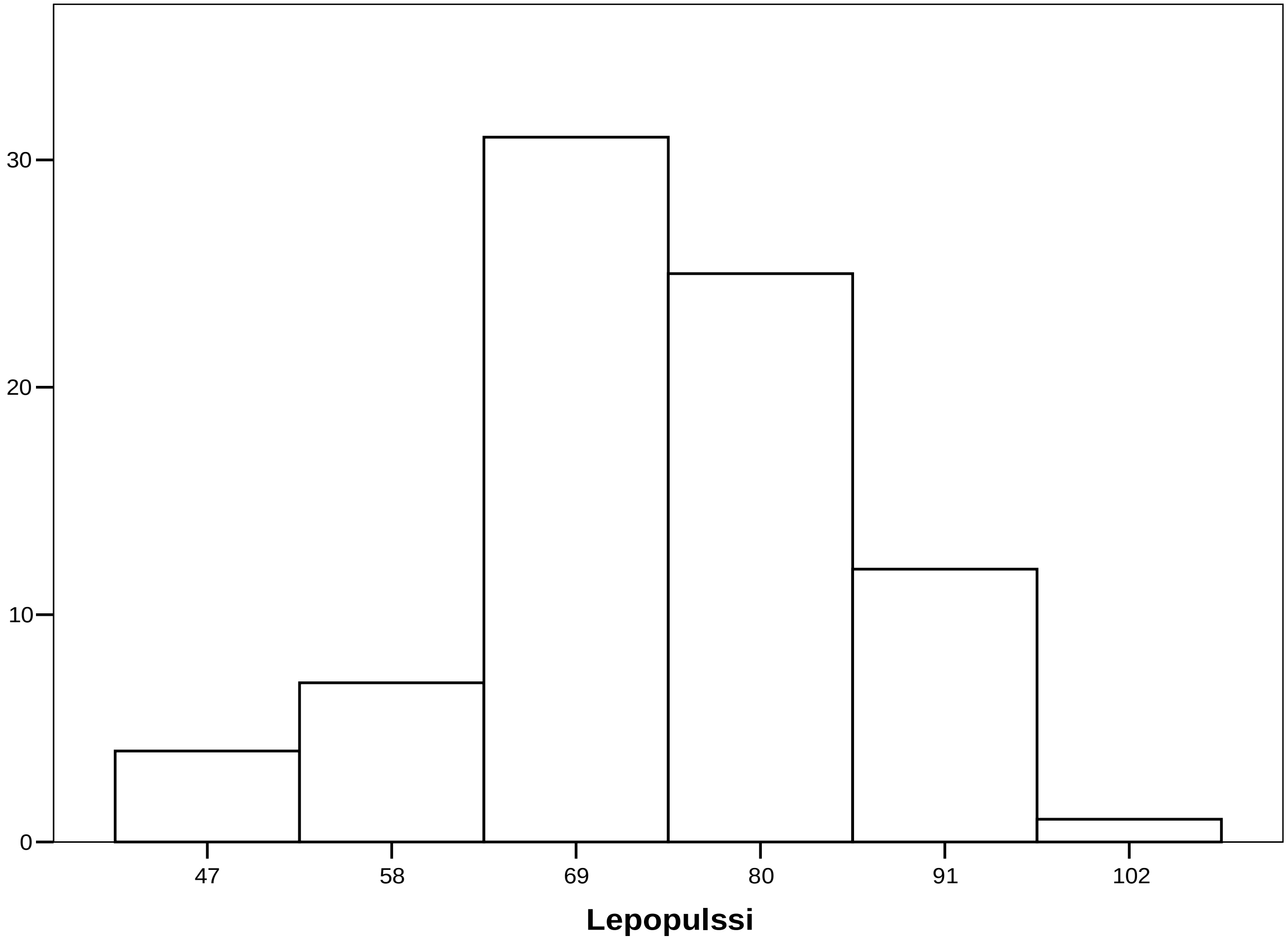
MTTTP1, luento 18.9.2018

KERTAUSTA

Esim. Pulssi-muuttujan frekvenssijakauma, aineisto luentomoniste liite 4

pyöristetyt luokkarajat	todelliset luokkarajat	luokka-keskus	frekvenssi
42–52	41,5–52,5	47	4
53–63	52,5–63,5	58	7
64–74	63,5–74,5	69	31
75–85	74,5–85,5	80	25
86–96	85,5–96,5	91	12
97–107	96,5–107,5	102	1

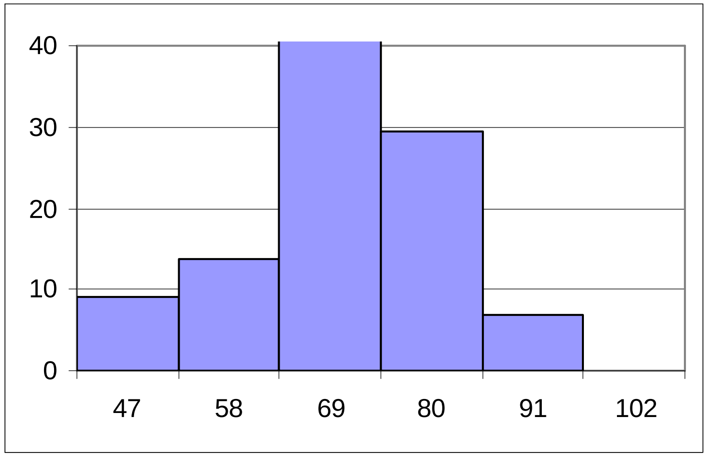
Graafinen esitys frekvenssihistogrammi



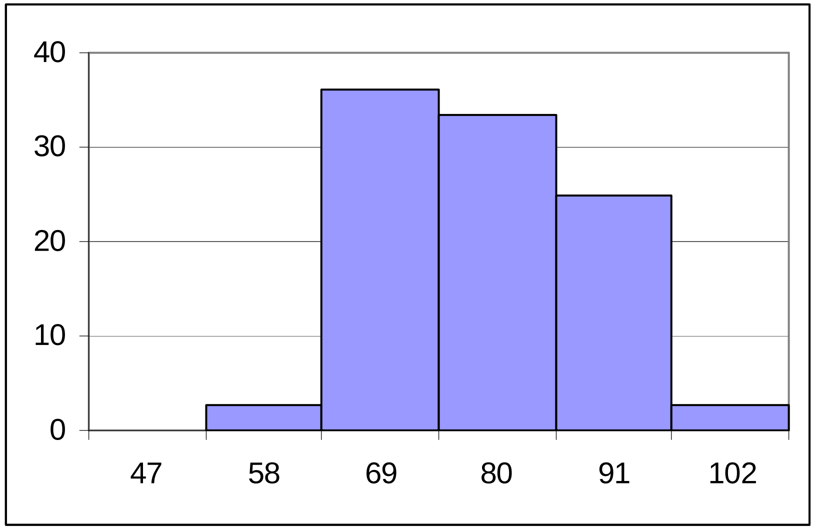
Huom. Piirretään todellisista luokkarajoista

Esim. 5.1.13. Pulssi-muuttujan frekvenssihistogrammit miehillä ja naisilla esimerkin 5.1.7 taulukosta, piirretään käyttäen prosentuaalisia frekvenssejä, jotta jakaumien vertailu olisi paremmin mahdollista.

Lepopulssin jakauma miehillä



Lepopulssin jakauma naisilla



Esim. Tampereelle 2009 myytyjä pieniä (alle 35 m^2) asuntoja, aineisto Tre_myydyt_asunnot_2009.sav sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

Muuttujat: Neliöt, Hinta, Rakennusvuosi, Sijainti, Kunto, Neliöhinta = Hinta/Neliöt

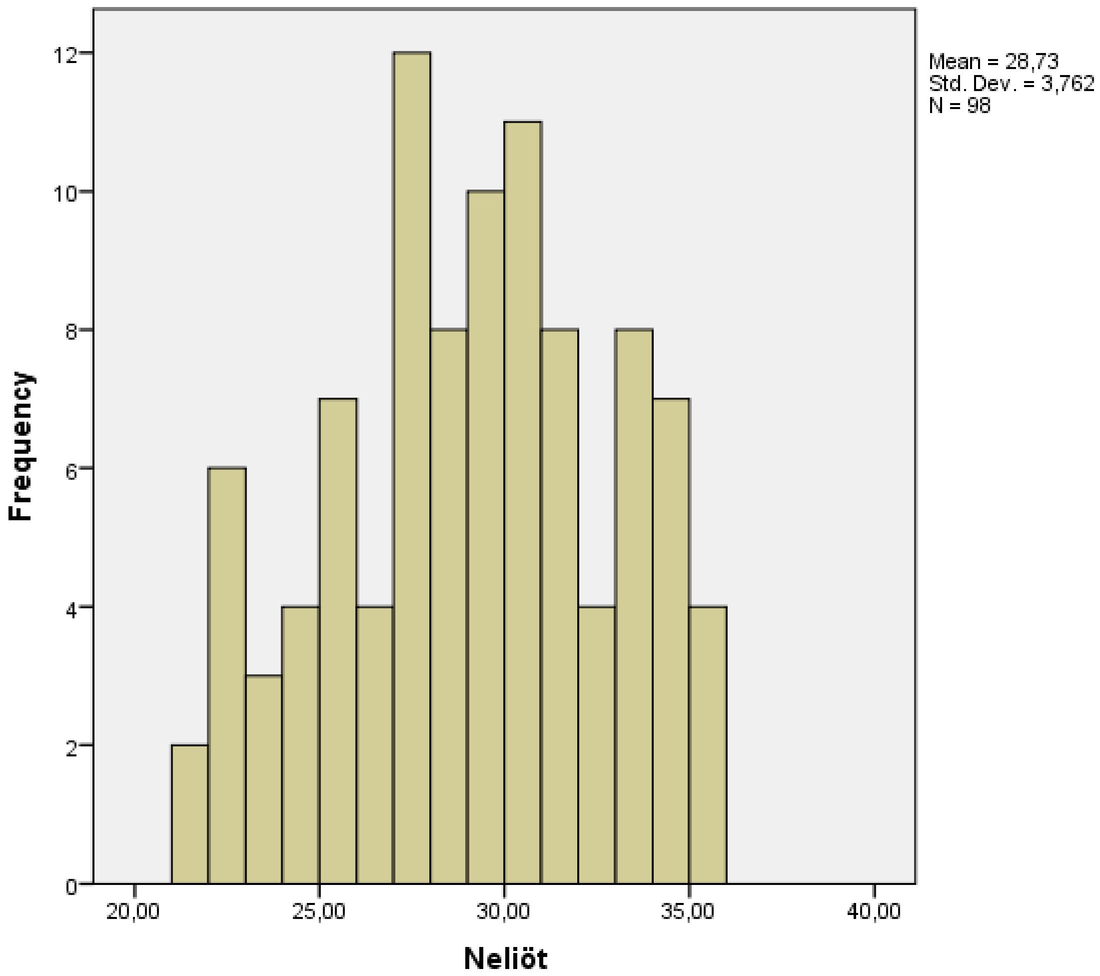
Jakaumia
(SPSS-ohjelman tuottamia taulukoita ja kuvia)

Sijainti

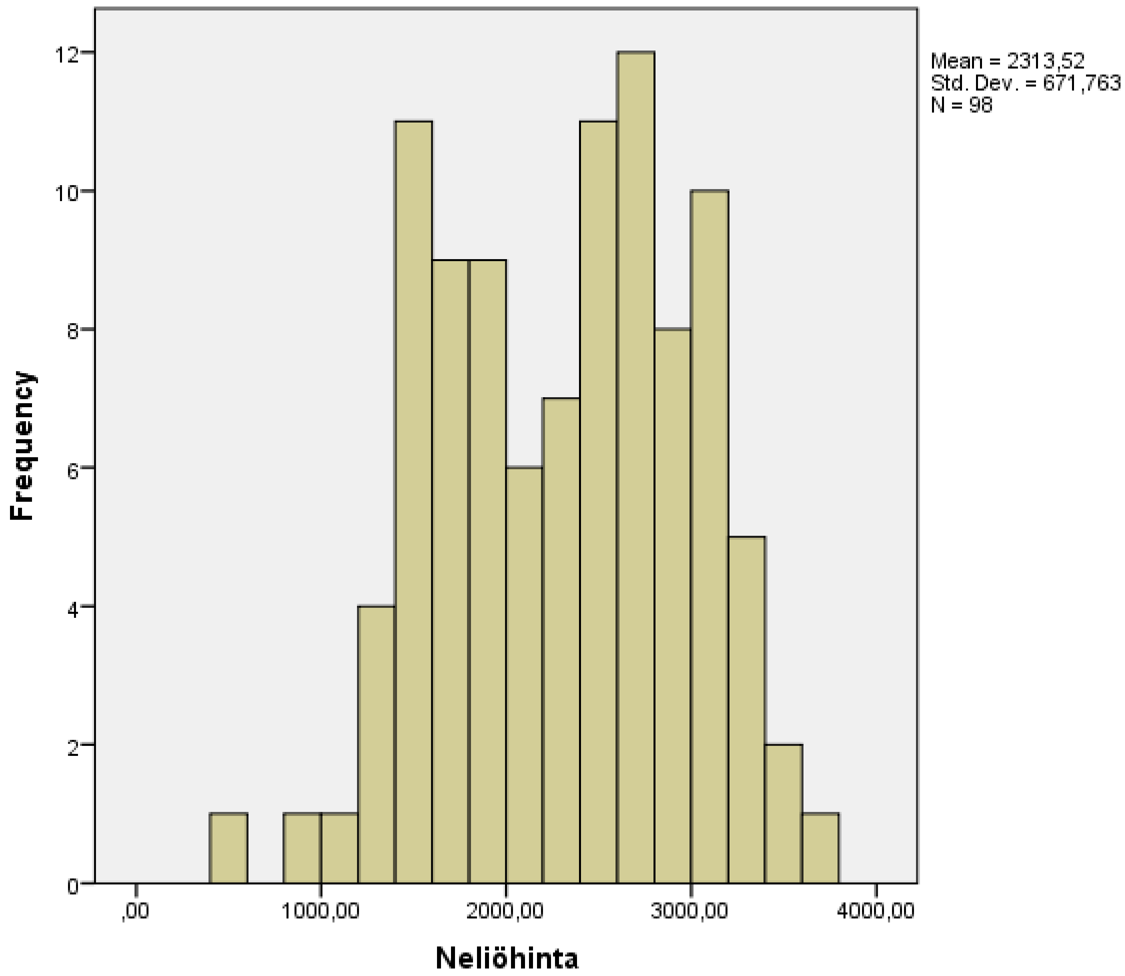
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Keskustassa	21	21,4	21,4	21,4
	Alle 5 km keskustasta	47	48,0	48,0	69,4
	Yli 5 km keskustasta	30	30,6	30,6	100,0
	Total	98	100,0	100,0	

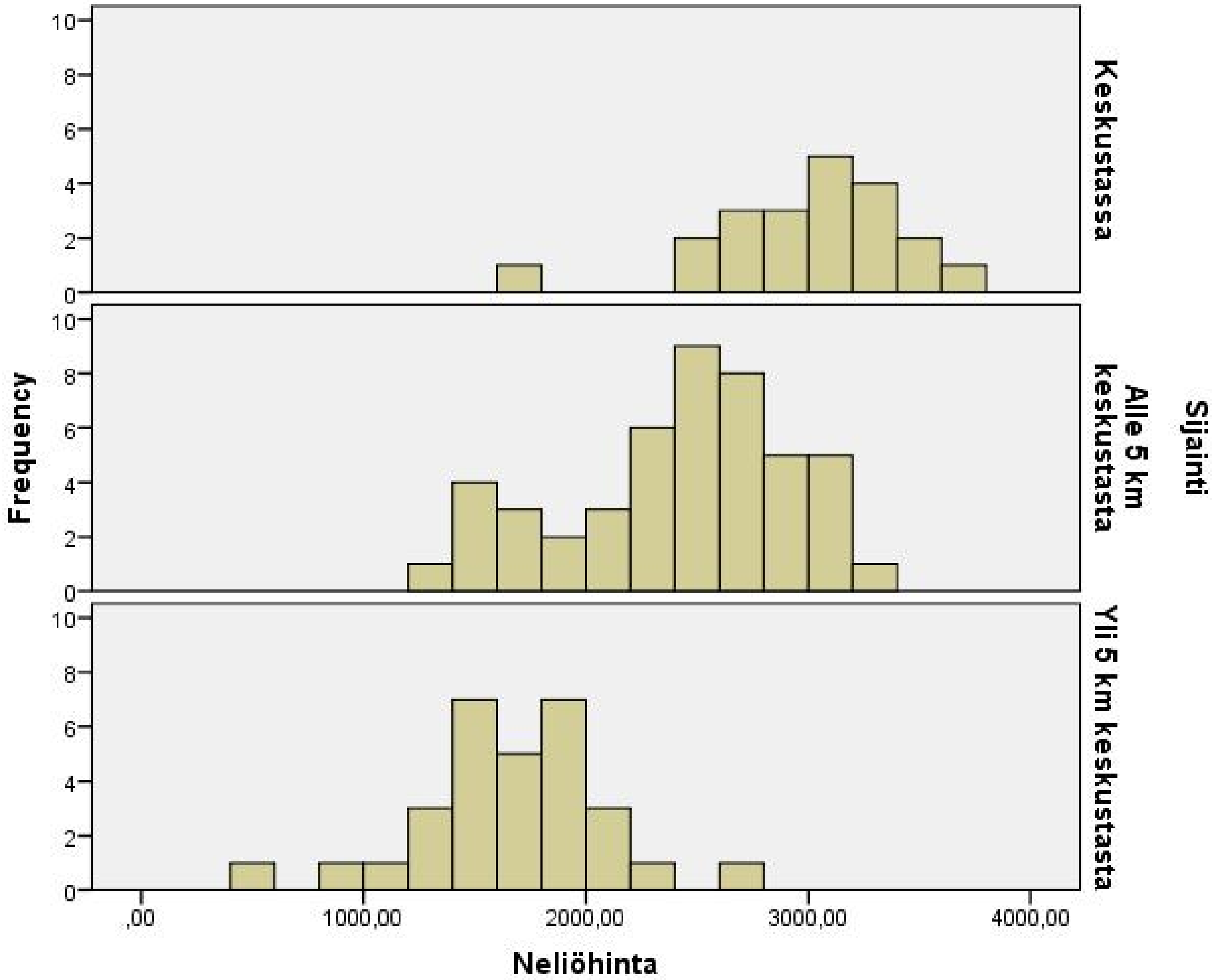
Kunto

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Hyvä	32	32,7	32,7	32,7
	Tyydyttävä	44	44,9	44,9	77,6
	Huono	22	22,4	22,4	100,0
	Total	98	100,0	100,0	



Mean = 28,73
Std. Dev. = 3,762
N = 98
Frequency
0
2
4
6
8
10
12
20,00
25,00
30,00
35,00
40,00
Neliöt


Mean = 2313,52
Std. Dev. = 671,763
N = 98
Frequency
12
10
8
6
4
2
0
,00
1000,00
2000,00
3000,00
4000,00
Neliöhinta



Frequency
10
8
6
4
2
0
Keskustassa
Alle 5 km keskustasta
Yli 5 km keskustasta
Sijainti
,00
1000,00
2000,00
3000,00
4000,00
Neliöhinta

5.1.3 Yksiulotteisen jakauman tunnuslukuja

Tunnusluvut

- kuvataan jakaumaa muuttujan arvoista lasketulla (tai arvojen avulla määritellyllä) luvulla
- kuvataan jakauman sijaintia, vaihtelua, vinoutta, huipukkuutta, jne.
- mitta-asteikko määrittää tunnusluvun valinnan

1) Sijainnin tunnuslukuja

Keskilukuja

- moodi (Mo)
- mediaani (Md), järjestysasteikollisuus
- keskiarvo, kvantitatiivisuus

Muita sijainnin tunnuslukuja

- ala- ja yläkvartiili, muut fraktiilit, järjestysasteikollisuus
- laatikko-jana –kuvio muodostetaan kvartiilien avulla

2) Vaihtelua mittaavia tunnuslukuja

- varianssi, keskihajonta, kvantitatiivisuus
- variaatiokerroin, suhdeasteikollisuus

3) Muita tunnuslukuja

- erilaisia vinous- ja huipukkuuskertoimia

1) Sijainnin tunnuslukuja

Keskilukuja

- Moodi (Mo) on se muuttujan arvo, joka esiintyy useimmin tai se luokka, jossa on eniten havaintoja

Esim. Lapsen sisarusten lukumäärä, esim. 5.1.29

Sisarusten lukumäärä	Frekv.
0	56
1	39
2	13
3	10
4	5
5	2
6	1
Yht.	126

Mo = 0

- Mediaani (*Md*) on sellainen muuttujan arvo, jota pienempiä ja suurempia arvoja on yhtä paljon. Muuttujan oltava vähintään järjestysasteikollinen.

Esim. 5.1.14. Tenttipisteet: 95, 86, 78, 90, 62, 73, 89

Md = 86

Esim. 5.1.29. Sisarusten lukumäärä

Md = 1

- Keskiarvo, kaava (1), http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/kaavat.pdf , vaaditaan kvantitatiivisuus

  Muuttujan x arvot tilastoyksiköittäin x_1, x_2, ..., x_n, tällöin

$$\boxed{\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

Esim. Etäisyydet, joista lepakot löysivät hyönteisiä, ks. Selityksiä ja esimerkkejä kaavoihin http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/esimerkit_kaavoihin.pdf

Esim. 5.1.15. Keskiarvo tenttipisteistä

(95+86+78+90+62+73+89)/7 = 81,9

Esim. 5.1.16. Lepopulssin keskiluvut, mediaani 74 ja keskiarvo 73,75.

SPSS-tulos:

Statistics

Pulssi

N	Valid	80
	Missing	0
Mean		73,7500
Median		74,0000
Std. Deviation		11,12814

Esim. 5.1.29. Sisarusten lukumäärän keskiarvo, keskiarvo frekvenssijakaumasta

Sisarusten lukumäärä	Frekv.
0	56
1	39
2	13
3	10
4	5
5	2
6	1
Yht.	126

$\bar{x} = (0 \cdot 56 + 1 \cdot 39 + 2 \cdot 13 + 3 \cdot 10 + 4 \cdot 5 + 5 \cdot 2 + 6 \cdot 1)/126 = 1{,}04$. Aineistossa lapsella on keskimäärin 1,04 sisarusta.

Muuttujan x keskistäminen

$x_i - \bar{x}$, i = 1, 2, ...,n

Keskiarvo ryhmäkeskiarvojen avulla

$\bar{x} = (n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k)/(n_1 + n_2 + \cdots + n_k)$

Esim. 5.1.20. Lepopulssin keskiarvo miehillä ja naisilla

Pulssi

Sukupuoli	Mean	Std. Deviation	N	Median
Mies	70,6364	11,27684	44	70,0000
Nainen	77,5556	9,80800	36	77,5000
Total	73,7500	11,12814	80	74,0000

73,7500 = (44·70,6364 + 36·77,5556)/80

Esim. 5.1.21. Voidaanko sadon määrää selittää käytetyllä viljelymenetelmällä?

satomäärä = selitettävä, riippuva muuttuja (y)
viljelymenetelmä = selittävä, riippumaton muuttuja (x)

Esim. 5.1.24. Voidaanko neliöhintaa selittää sijainnilla?

y = neliöhinta
x = sijainti

Neliöhinta

Sijainti	Mean	N	Std. Deviation
Hervanta	1752,6063	127	456,78817
Kaleva	2569,1688	77	394,44750
Keskusta	3118,4512	164	712,47930
Tesoma	1593,3333	54	484,13026
Total	2412,0213	422	855,94477

Esim. 5.1.17. Keskiluvut symmetristen ja vinojen jakaumien tapauksessa,
http://www.sis.uta.fi/tilasto/tiltp7/moniste_4.pdf

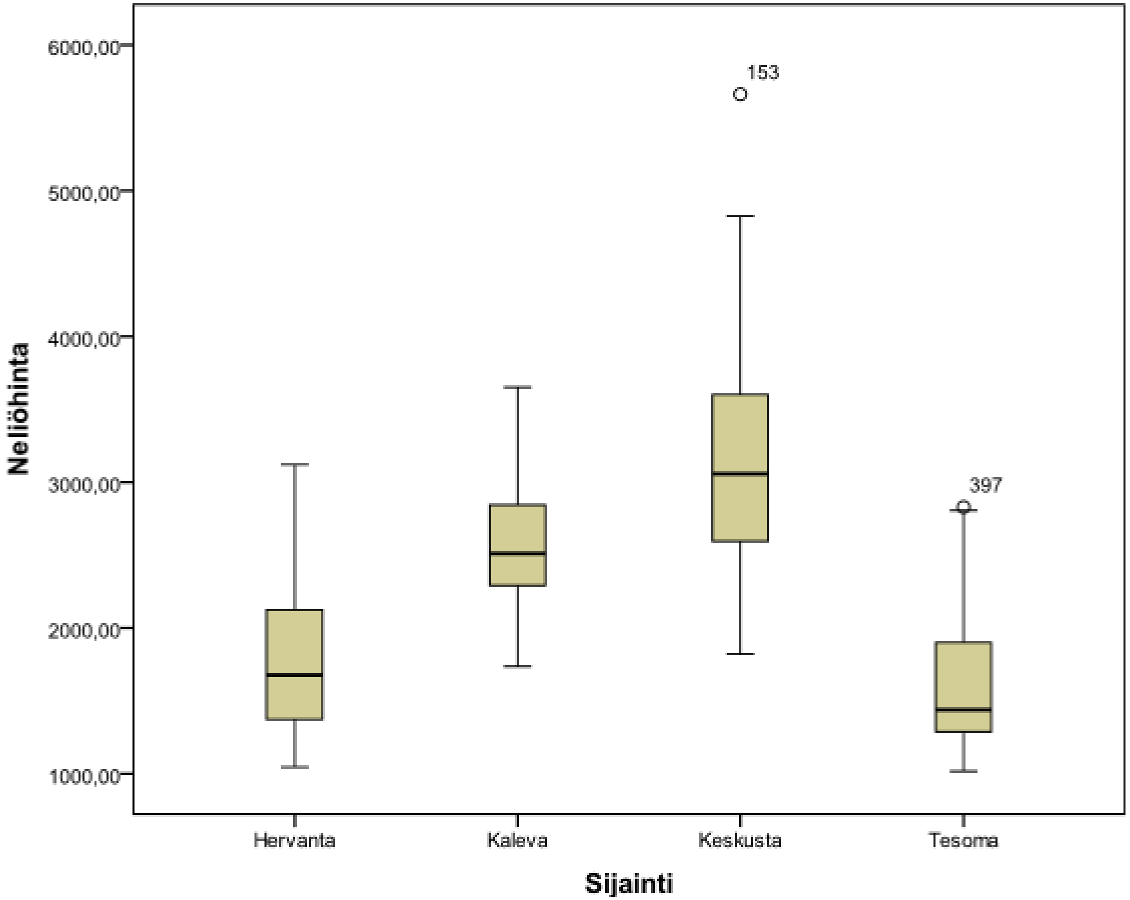
Muita sijainnin tunnuslukuja

- Ala- ja yläkvartiili, muut fraktiilit

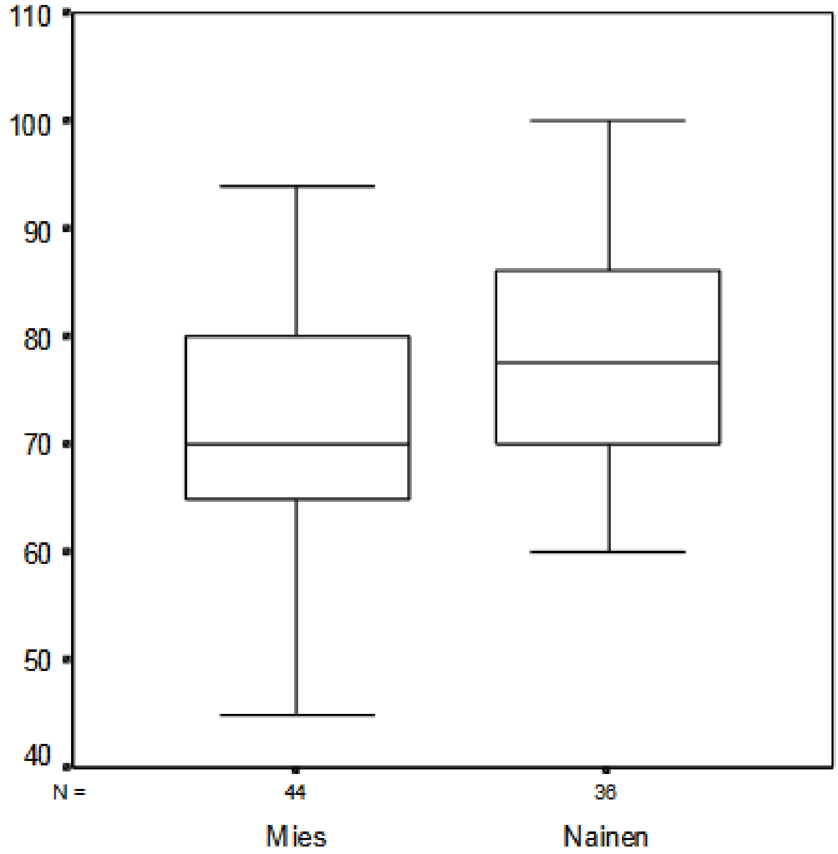
http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=29

Laatikko-jana –kuvio muodostetaan kvartiilien avulla

Esim. 5.1.25. Neliöhinta sijainnin mukaan



Esim. 5.1.26. Lepopulssi miehillä ja naisilla



Sivut 21 – 22 seuraavalle luennolle

2) Vaihtelua mittaavia tunnuslukuja
http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=32

- Varianssi, kaava (2)
  http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/kaavat.pdf

$$\boxed{s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Mittaa muuttujan arvojen keskittymistä keskiarvon ympärille, sallittu kvantitatiivisen muuttujan yhteydessä.

- Keskihajonta, kaava (3)

$$\boxed{s_x = \sqrt{s_x^2}}$$

Esim. Etäisyydet, joista lepakot löysivät hyönteisiä, ks. Selityksiä ja esimerkkejä kaavoihin http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/esimerkit_kaavoihin.pdf

Esim. 5.1.28. Otosvarianssin laskeminen tenttipisteistä 95, 86, 78, 90, 62, 73, 89

$s^2 = ((95-81{,}9)^2 + (86-81{,}9)^2 + \ldots + (89-81{,}9)^2) / (7-1) = 132{,}5$

$s = 11{,}5$.

Esim. 5.1.35. Normaalijakauma http://www.sis.uta.fi/tilasto/tiltp7/moniste_6.pdf

Esim. Laskuri http://vassarstats.net/, jossa keskiarvon ja varianssin lasku http://vassarstats.net/vsmisc.html

MTTTP1, luento 20.9.2018

KERTAUSTA JA TÄYDENNYSTÄ

Tunnusluvut

1) Sijainnin tunnuslukuja

Keskilukuja
- moodi (Mo)
- mediaani (Md)
- keskiarvo, kaava (1)

Muita sijainnin tunnuslukuja
- ala- ja yläkvartiili, muut fraktiilit

2) Vaihtelua mittaavia tunnuslukuja

http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=32

- Varianssi, kaava (2)
  http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/kaavat.pdf

$$\boxed{s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

  Mittaa muuttujan arvojen keskittymistä keskiarvon ympärille, sallittu kvantitatiivisen muuttujan yhteydessä.

- Keskihajonta, kaava (3)

$$\boxed{s_x = \sqrt{s_x^2}}$$

Esim. Etäisyydet, joista lepakot löysivät hyönteisiä, ks. Selityksiä ja esimerkkejä kaavoihin http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/esimerkit_kaavoihin.pdf

Esim. 5.1.28. Otosvarianssin laskeminen tenttipisteistä
95, 86, 78, 90, 62, 73, 89

$s^2 = ((95-81{,}9)^2 + (86-81{,}9)^2 + \ldots + (89-81{,}9)^2) / (7-1) = 132{,}5$

$s = 11{,}5.$

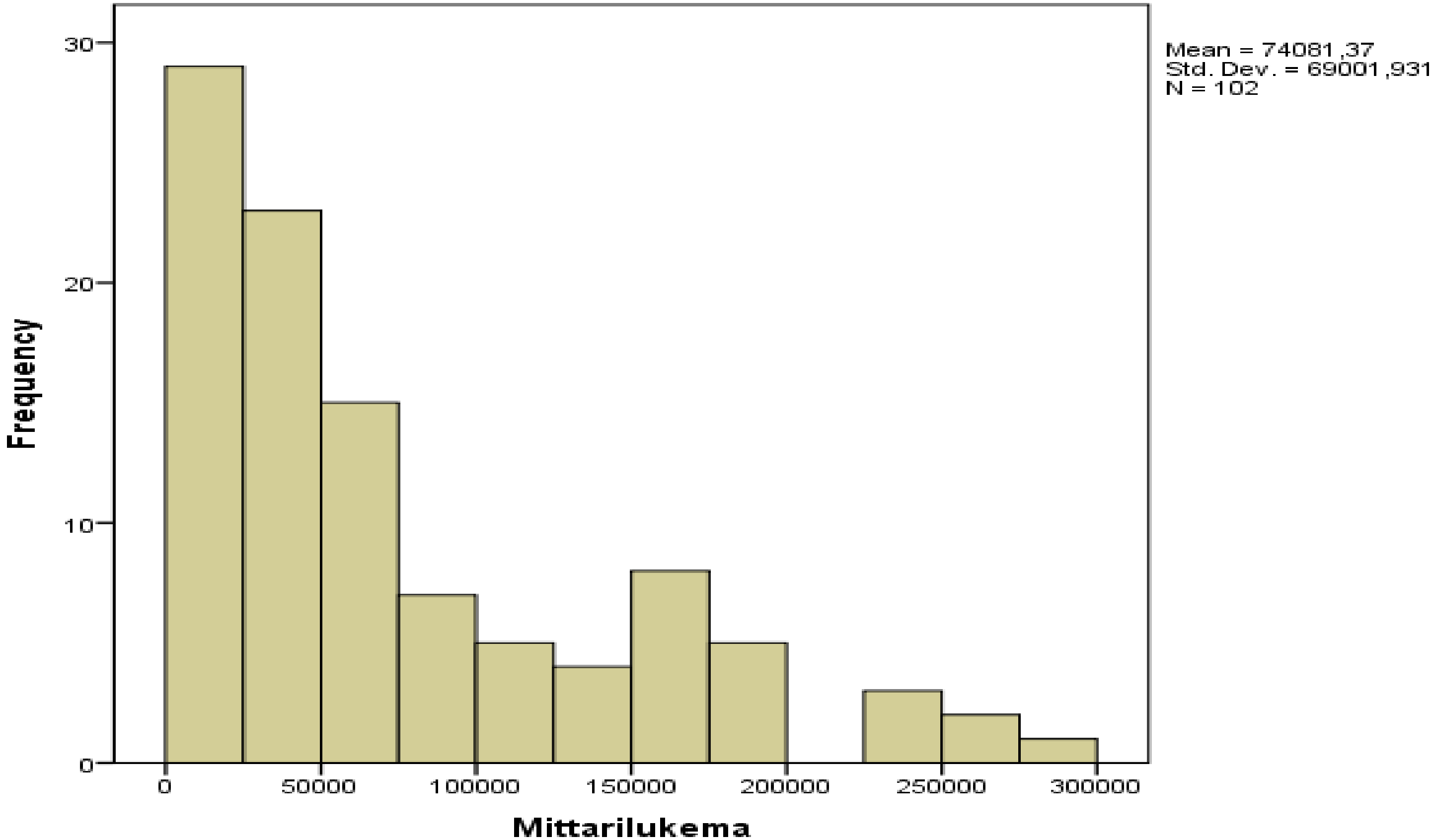
Esim. 5.1.35. Normaalijakauma
http://www.sis.uta.fi/tilasto/tiltp7/moniste_6.pdf

Esim. Laskuri http://vassarstats.net/, jossa keskiarvon ja varianssin lasku http://vassarstats.net/vsmisc.html

3) Muita tunnuslukuja

Voidaan mitata esim. jakauman vinoutta ja huipukkuutta

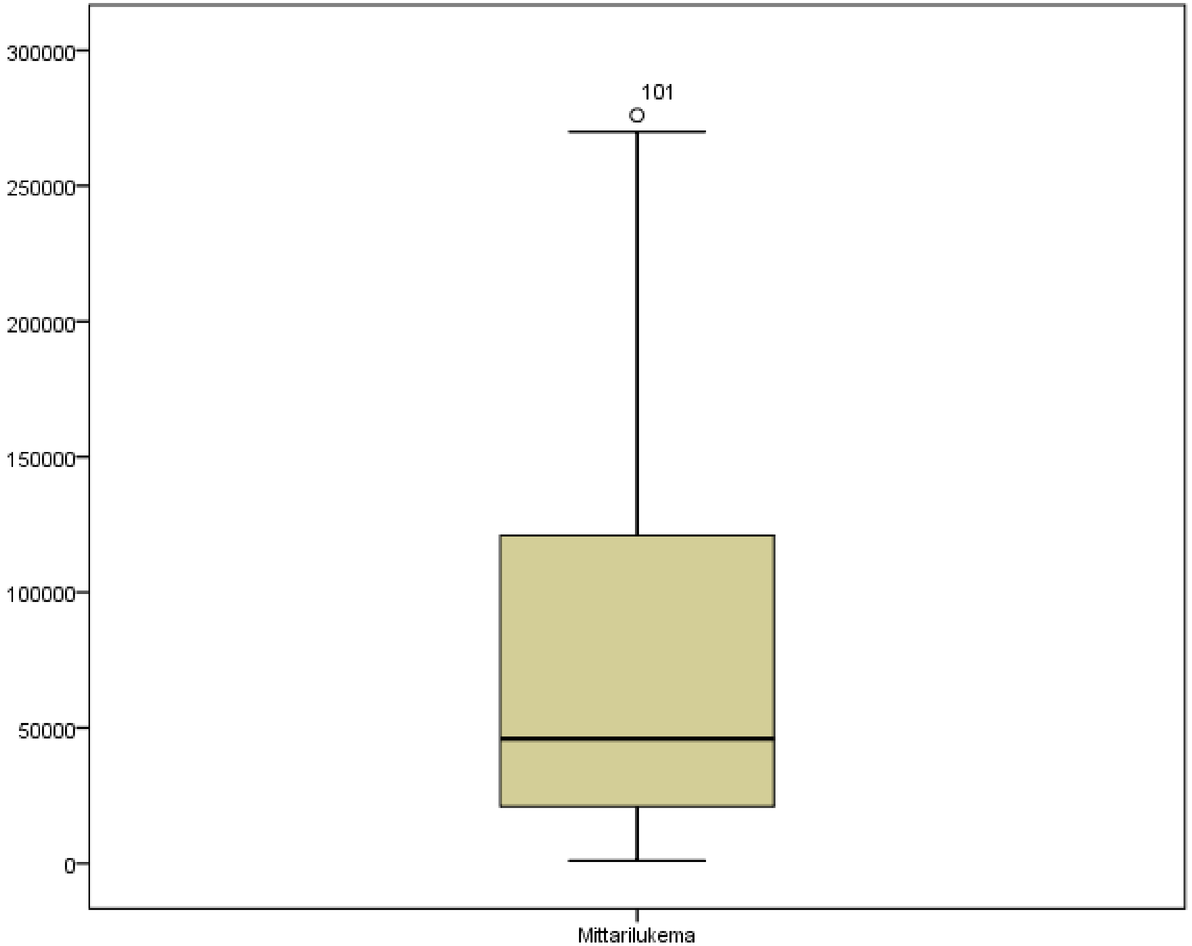
Esim.
Sivulta _
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/
aineisto Toyota.sav, jossa Toyota Avensis -farmariautoja vuosilta 2007 - 2009, oikotie.fi -sivustolta 2.2.2010.



Statistics

Mittarilukema

N	Valid	102
	Missing	0
Mean (keskiarvo)		74081,37
Median (mediaani)		46000,00
Std. Deviation (keskihajonta)		69001,931
Percentiles	25	21000,00
	50	46000,00
	75	121250,00


300000
250000
200000
150000
100000
50000
0
101
Mittarilukema

Moottorin tilavuus

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1,6	5	4,9	4,9	4,9
	1,8	25	24,5	24,5	29,4
	2,0	55	53,9	53,9	83,3
	2,2	17	16,7	16,7	100,0
	Total	102	100,0	100,0	

Taulukosta laskettuna:

Md = 2,0,

Mo = 2,0,

Keskiarvo

$$(1{,}6 \cdot 5 + 1{,}8 \cdot 25 + 2{,}0 \cdot 55 + 2{,}2 \cdot 17)/102 = 1{,}965$$

Statistics

Moottorin tilavuus

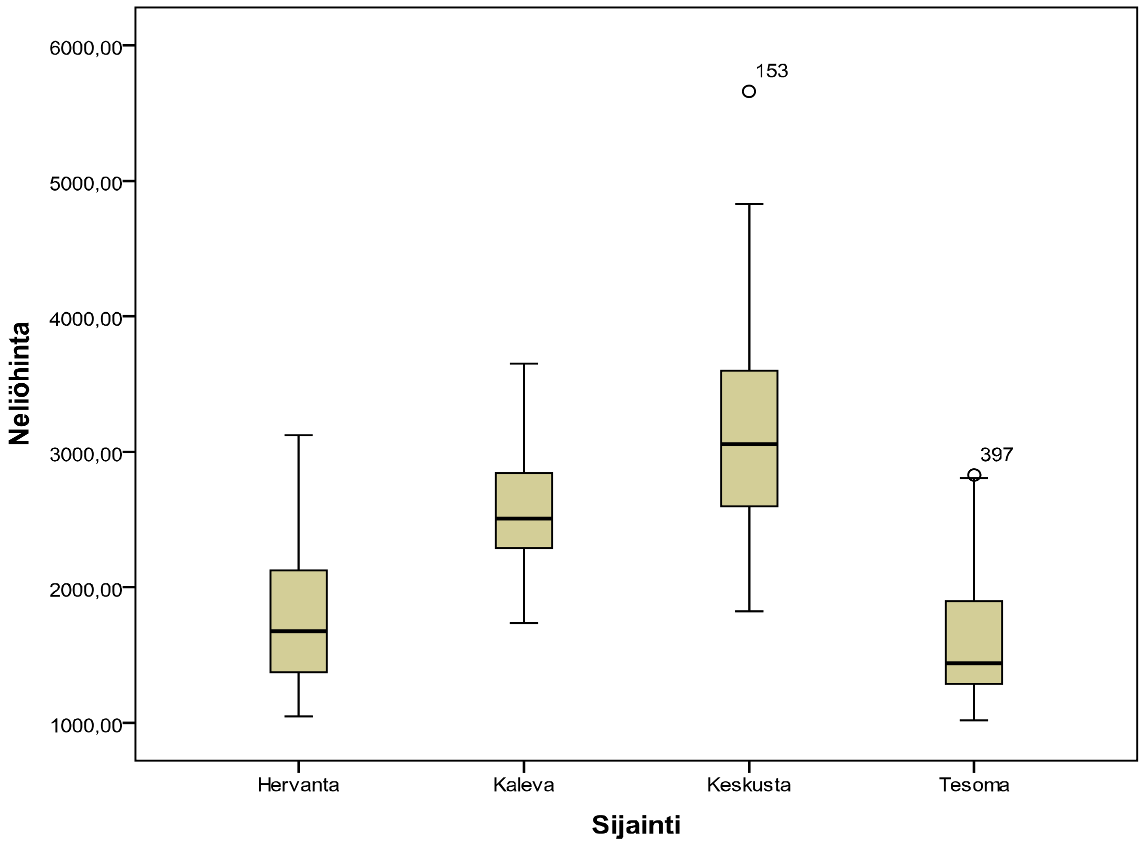
N	Valid	102
	Missing	0
Mean		1,965
Median		2,000
Std. Deviation		,1526
Percentiles	25	1,800
	50	2,000
	75	2,000

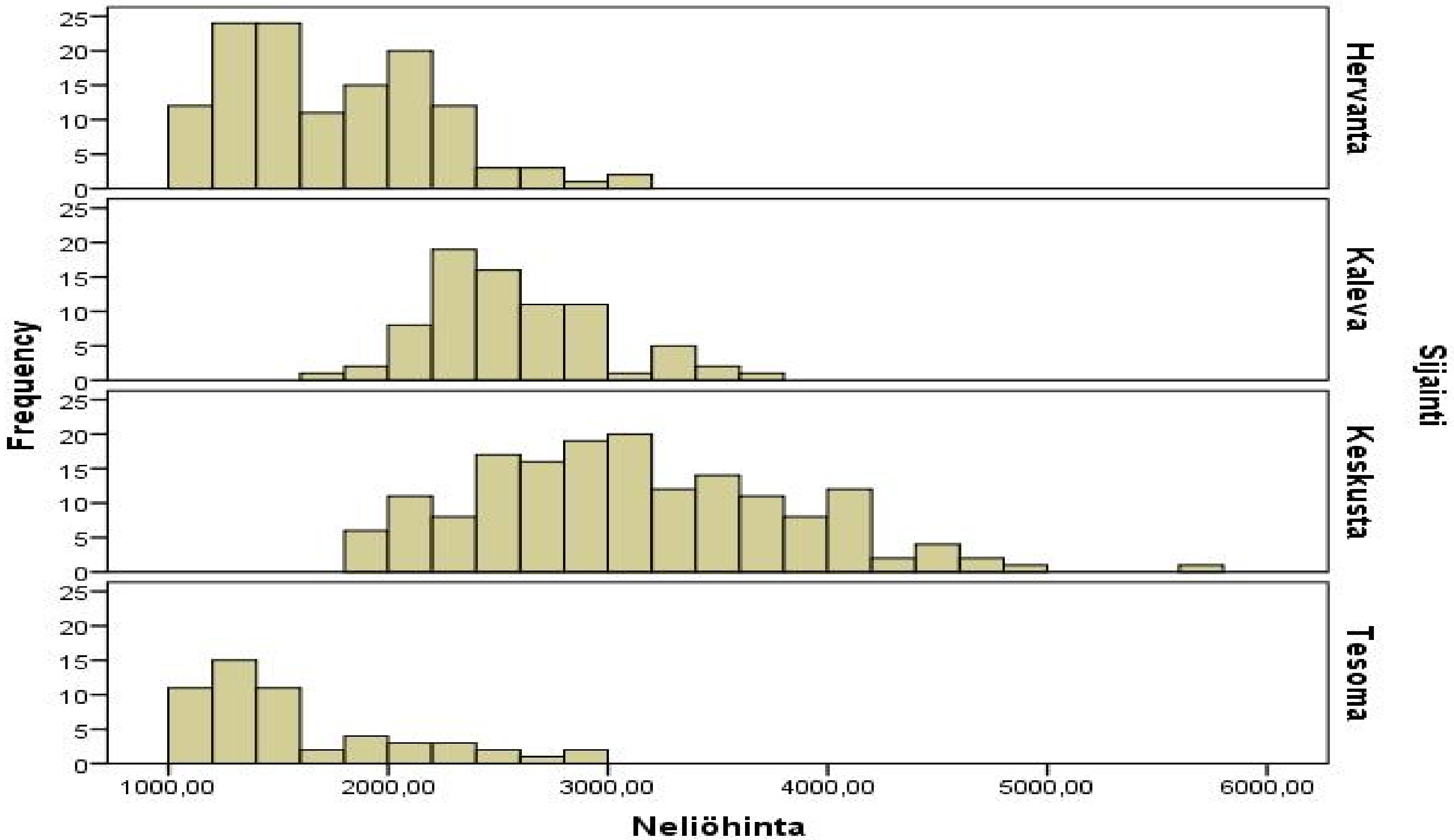
Väri

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Harmaa	10	9,8	12,7	12,7
	Hopea	43	42,2	54,4	67,1
	Musta	15	14,7	19,0	86,1
	Punainen	5	4,9	6,3	92,4
	Sininen	2	2,0	2,5	94,9
	Ruskea	4	3,9	5,1	100,0
	Total	79	77,5	100,0	
Missing	System	23	22,5		
Total		102	100,0		

Hopeisia on eniten (54,4 %).

Esim. 5.1.24, 5.1.25, 5.1.27
Myytyjen kerrostaloasuntojen neliöhintoja Tampereella, sivun
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/
aineisto Tre_myydyt_asunnot_2012.sav





	Hervanta	Kaleva	Keskusta	Tesoma
Keskiarvo	1753	2569	3118	1593
Mediaani	1677	2510	3058	1438
Keskihajonta	457	394	712	484

Esim. 5.1.30. Lisäaineen vaikutus teräksen kovuusindeksiin

Tuote-erä	1	2	3	4	5	6	7	8	9	10
Lisäaine A	22	26	29	22	31	34	31	20	33	34
Lisäaine B	27	25	31	27	29	41	32	27	32	34
Erotus	-5	1	-2	-5	2	-7	-1	-7	1	0

Laskurilla http://vassarstats.net/vsmisc.html erotuksen keskiarvon ja varianssin lasku

Lineaarinen muunnos muuttujalle x

$$y_i = ax_i + b, \; i = 1, 2, ..., n$$

- vaikutus keskiarvoon

$$\bar{y} = a\bar{x} + b$$

mittayksikkö vaikuttaa keskiarvon

- vaikutus keskihajontaan

$$s_y = |a| s_x$$

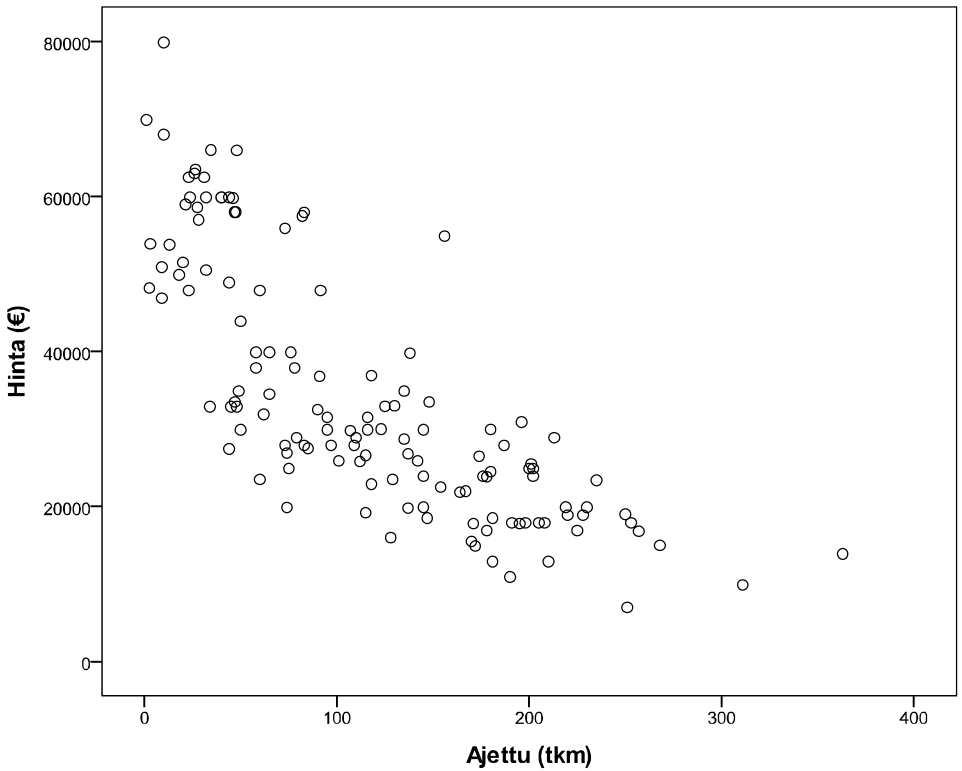
mittayksikkö vaikuttaa keskihajontaan

Muuttujan x standardointi

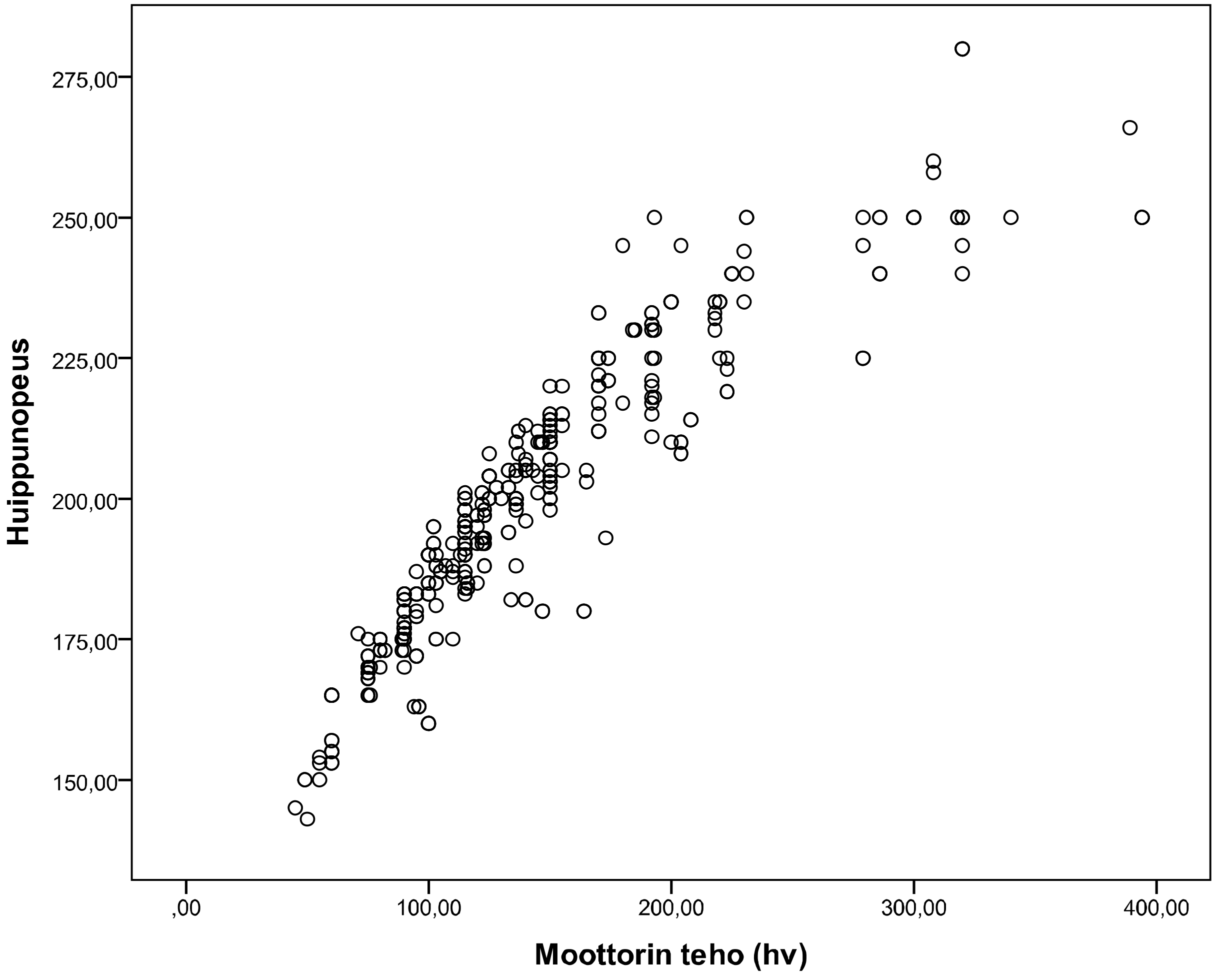
$$z_i = \frac{x_i - \bar{x}}{s}$$

5.2 Kaksiulotteinen jakauma
5.2.1 Pisteparvi

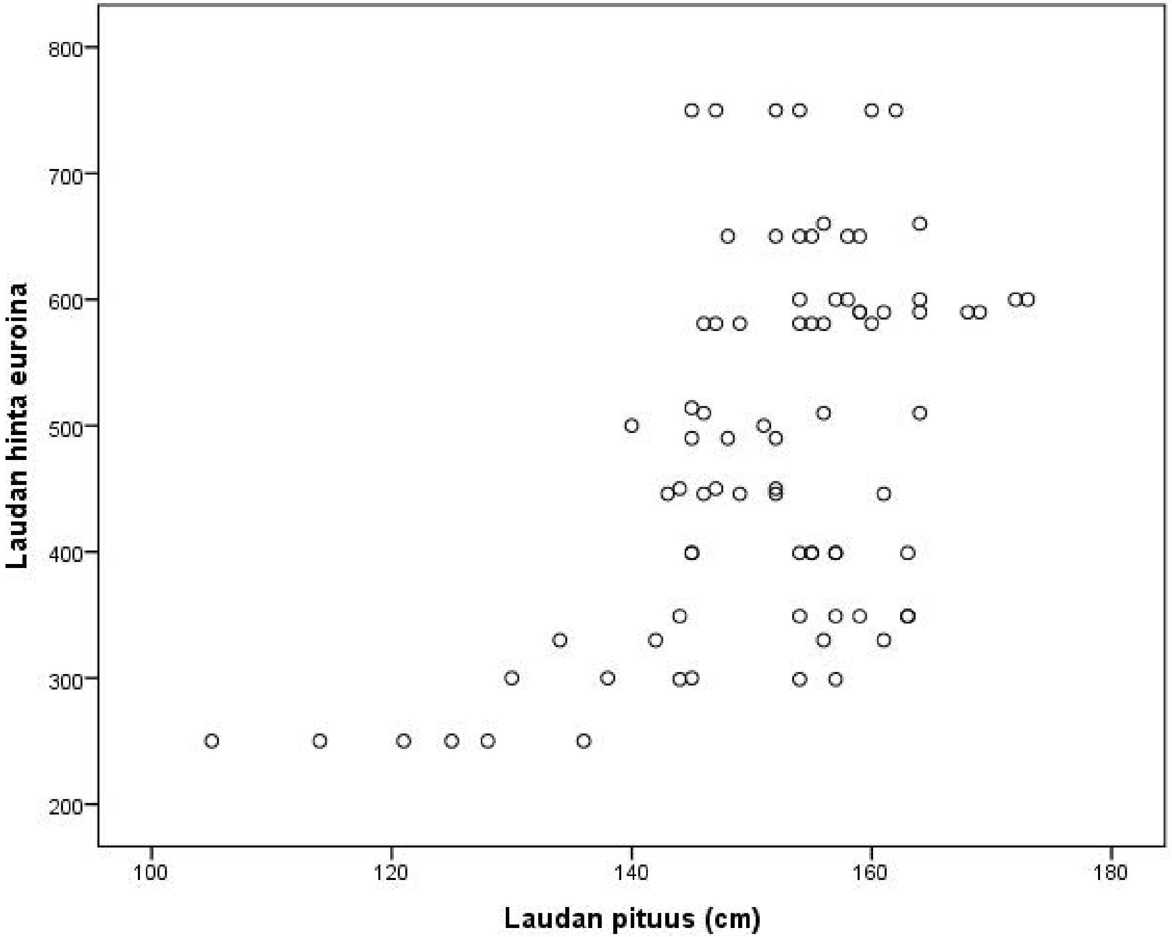
Esim. Auton hinta ja ajetut kilometrit, aineistona Audi A6 –henkilöautoja sivulta https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/



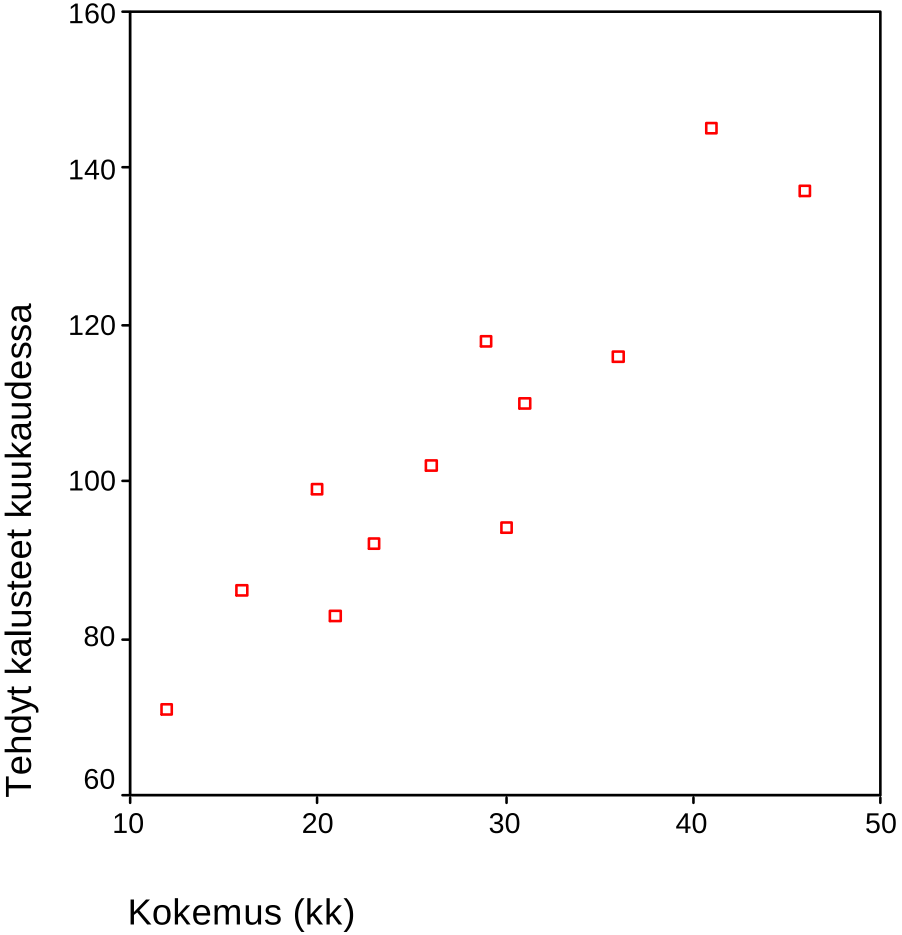
Esim. Auton huippunopeus ja teho, aineistona auto94.sav (mikroluokissa)



Esim. Lumilaudan hinta ja pituus, aineistona Lumilaudat.sav sivulta https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/



Esim. 5.2.2. Tehdyt kalusteet ja työntekijän kokemus,



Kaksiulotteisessa jakaumassa tarkastellaan kahta muuttujaa samanaikaisesti. Tutkitaan muuttujien välisiä riippuvuussuhteita.

Pisteparvi on graafinen esitys, jos selitettävä muuttuja kvantitatiivinen.

5.2.2 Ristiintaulukko

Esim. Miesten, naisten ja lasten lumilaudat valmistusmaittain, aineistona Lumilaudat.sav sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

Kenelle lauta on tarkoitettu * Merkki luokiteltuna maan mukaan Crosstabulation

Count

		Merkki luokiteltuna maan mukaan		
		Kanada	USA	Total
Kenelle lauta on tarkoitettu	Miehet	22	29	51
	Naiset	11	7	18
	Lapset	7	4	11
Total		40	40	80

Esim. 5.2.5. Automallien koot valmistusmaittain

		Valmistusmaa			
		USA	Eur.	Japani	
	Iso	36	4	2	42
Koko	Kesk.	53	17	54	124
	Pieni	26	19	92	137
		115	40	148	303

Koko-muuttujan ehdolliset prosenttijakaumat eli koko-muuttujan prosentuaaliset jakaumat erikseen valmistusmaittain:

		Valmistusmaa		
		USA	Eur.	Japani
	Iso	31,30	10,00	1,35
Koko	Kesk.	46,09	42,50	36,49
	Pieni	22,61	47,50	62,16
		100,00	100,00	100,00

Esim. 5.2.6. Markkinointisuunnitelma tavaratalon koon mukaan

		Suunnitelma on	ei	yht.
Henkilöstö-	alle 100	13	10	23
määrä	100 – 500	18	12	30
	yli 500	32	6	38

Markkinointisuunnitelman olemassaolon ehdolliset prosenttijakaumat (koon mukaan)

		Suunnitelma on	ei
Henkilöstö-	alle 100	56,6	43,5
määrä	100 – 500	60,0	40,0
	yli 500	84,2	15,8

Esim. Miesten, naisten ja lasten lumilaudat valmistusmaittain, aineistona Lumilaudat.sav sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

Kenelle lauta on tarkoitettu * Merkki luokiteltuna maan mukaan Crosstabulation

% within Merkki luokiteltuna maan mukaan

		Merkki luokiteltuna maan mukaan		
		Kanada	USA	Total
Kenelle lauta on tarkoitettu	Miehet	55,0%	72,5%	63,7%
	Naiset	27,5%	17,5%	22,5%
	Lapset	17,5%	10,0%	13,8%
Total		100,0%	100,0%	100,0%

Esim. Asunnon kunto sijainnin mukaan, aineistona Tre_myydyt_asunnot_2010 sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

Huoneiston kunto * Sijainti Crosstabulation

Count

		Sijainti				
		Keskusta	Kaleva	Hervanta	Härmälä	Total
Huoneiston kunto	huono	3	1	1	0	5
	hyvä	102	31	91	60	284
	tyydyttävä	39	21	37	7	104
Total		144	53	129	67	393

Onko tässä korjattavaa?

Ristiintaulukko yleisesti

				x		
		E_1	E_2		E_J	
	F_1	f_{11}	f_{12}	...	f_{1J}	$f_{1\cdot}$
	F_2	f_{21}	f_{22}	...	f_{2J}	$f_{2\cdot}$
y	.	.	.		.	.
	.	.	.		.	.
	.	.	.		.	.
	F_I	f_{I1}	f_{I2}	...	f_{IJ}	$f_{I\cdot}$
		$f_{\cdot 1}$	$f_{\cdot 2}$	...	$f_{\cdot J}$	$f_{\cdot\cdot} = n$

Tutkitaan y:n riippuvuutta x:stä vertaamalla y:n ehdollisia prosenttijakaumia.

6 AIKASARJOISTA

Määritelmä
http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=51

Graafinen esitys
http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=51

Esimerkkejä luentomonisteen esimerkeissä 6.1.1.- 6.1.6.

Harjoitustyön riippuvuustarkastelut

Ks. http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/htyop118.pdf#page=4

Riippuvuustarkastelu 1

Tarkastelu, jossa y (selitettävä) on kvantitatiivinen ja x (selittäjä) kvalitatiivinen. Menetelminä laatikko-jana-kuvio, ryhmäkeskiarvot, muut tarvittavat tunnusluvut sekä päättely riippumattomien otosten t-testi avulla (käytetään näitä jokaista). Jos ryhmiä enemmän kuin kaksi, niin t-testissä voi tarkastella niistä kahta tai yhdistellä ryhmiä sopivasti.

Ks. luentomonisteen esimerkit 5.1.25, 5.1.26, 7.7.9, 7.7.11.

Riippuvuustarkastelu 2

Tarkastelu, jossa sekä y että x kvalitatiivisia (kvantitatiiviset voi myös luokitella), selitettävä muuttuja eri kuin riippuvuustarkastelussa 1. Menetelmänä ristiintaulukko ja χ^2–riippumattomuustesti.

Ks. luentomonisteen esimerkit 7.7.6, 7.7.7, 7.7.8.

MTTTP1, luento 25.9.2018

KERTAUSTA

- Varianssi, kaava (2)
  http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/kaavat.pdf

$$\boxed{s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}$$

Mittaa muuttujan arvojen keskittymistä keskiarvon ympärille, sallittu kvantitatiivisen muuttujan yhteydessä.

Esim. 5.1.30. Lisäaineen vaikutus teräksen kovuusindeksiin

Tuote-erä	1	2	3	4	5	6	7	8	9	10
Lisäaine A	22	26	29	22	31	34	31	20	33	34
Lisäaine B	27	25	31	27	29	41	32	27	32	34
Erotus	-5	1	-2	-5	2	-7	-1	-7	1	0

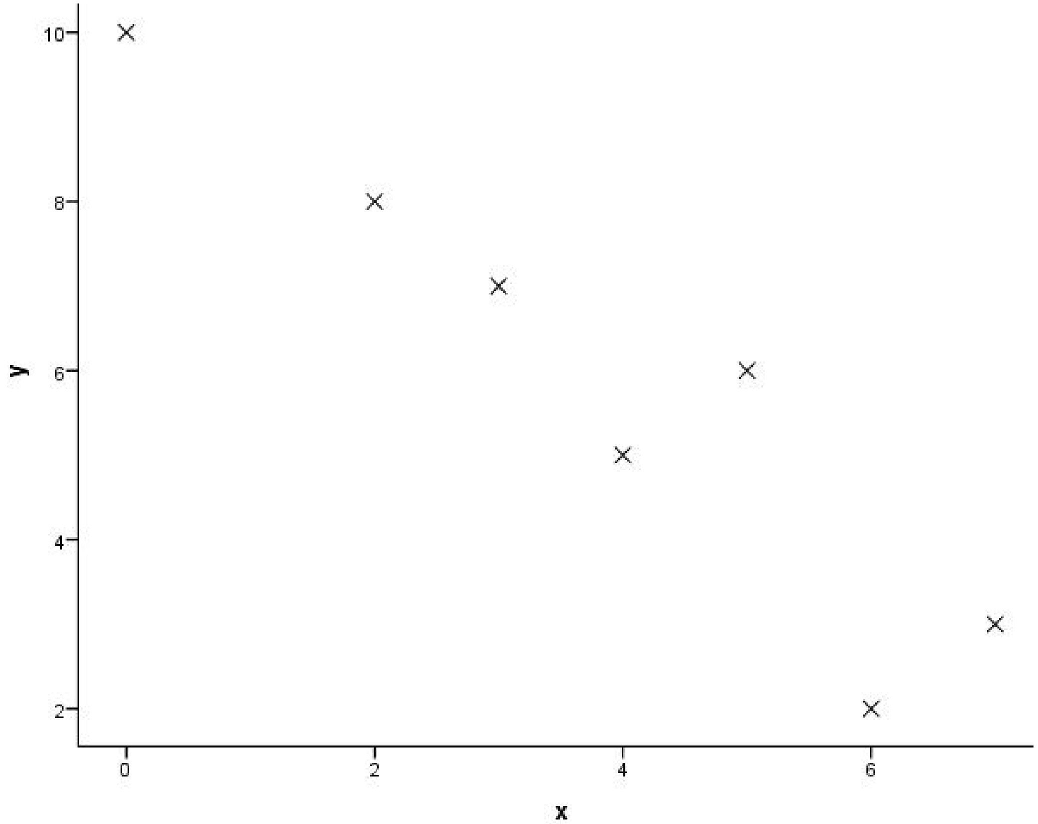
$\bar{x} = (-5+1-2-5+2-7-1-7+1+0)/10 = -2,3$

Lisäaineiden vaikutuksessa teräksen kovuuteen ei eroja, jos erotuksen keskiarvo riittävän lähellä nollaa. Päättely testauksen avulla.

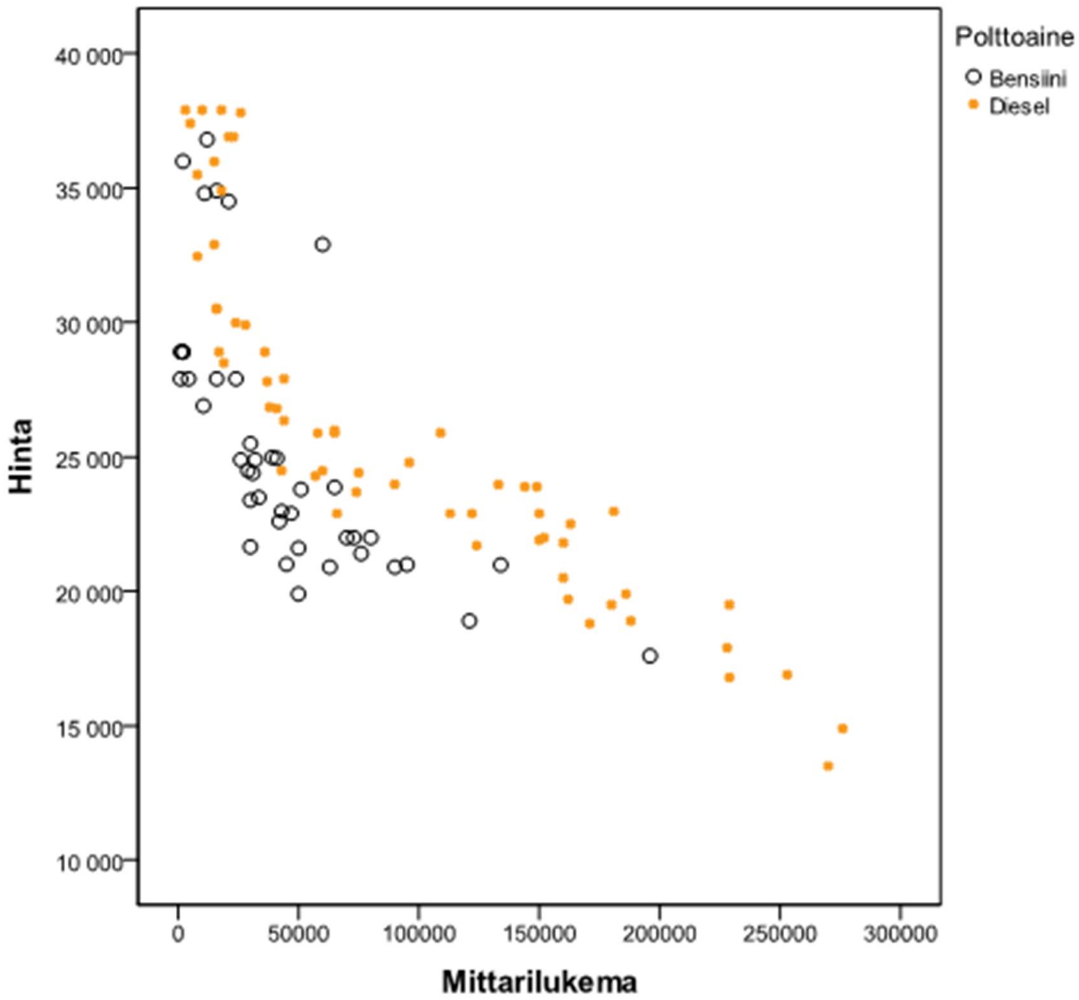
$s^2 = ((-5+2,3)^2 + (1+2,3)^2 + \ldots + (0+2,3)^2)/(10-1) =$
$11,79$
$s = 3,4.$

- Kaksiulotteinen jakauma

Pisteparvi, graafinen esitys



Esim. Toyota Avensis –farmariautoja



Ristiintaulukko

Esim. Toyota Avensis –farmariautoja, nelikenttä (2x2-taulukko)

		Vaihteisto		Total
		Automaatti	Manuaali	
Polttoaine	Bensiini	12	29	41
	Diesel	7	51	58
Total		19	80	99

Esim. Asunnon kunto sijainnin mukaan, aineistona Tre_myydyt_asunnot_2010 sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

		Sijainti				
		Keskusta	Kaleva	Hervanta	Härmälä	Total
Huoneiston kunto	huono	3	1	1	0	5
	hyvä	102	31	91	60	284
	tyydyttävä	39	21	37	7	104
Total		144	53	129	67	393

		Sijainti				
		Keskusta	Kaleva	Hervanta	Härmälä	Total
Huoneiston kunto	Huono tai tyydyttävä	29,2%	41,5%	29,5%	10,4%	27,7%
	Hyvä	70,8%	58,5%	70,5%	89,6%	72,3%
Total		100,0%	100,0%	100,0%	100,0%	100,0%

On eroja, $p = 0{,}002$

5.2.3 Kaksiulotteisen jakauman tunnuslukuja

- Mitataan kahden muuttujan välistä riippuvuuden voimakkuutta
- Ristiintaulukosta kontingenssikerroin
- Kvantitatiivisista muuttujista lineaarisen riippuvuuden voimakkuuden mittari korrelaatiokerroin (r)
- Järjestysasteikollisilla muuttujilla järjestyskorrelaatiokertoimet

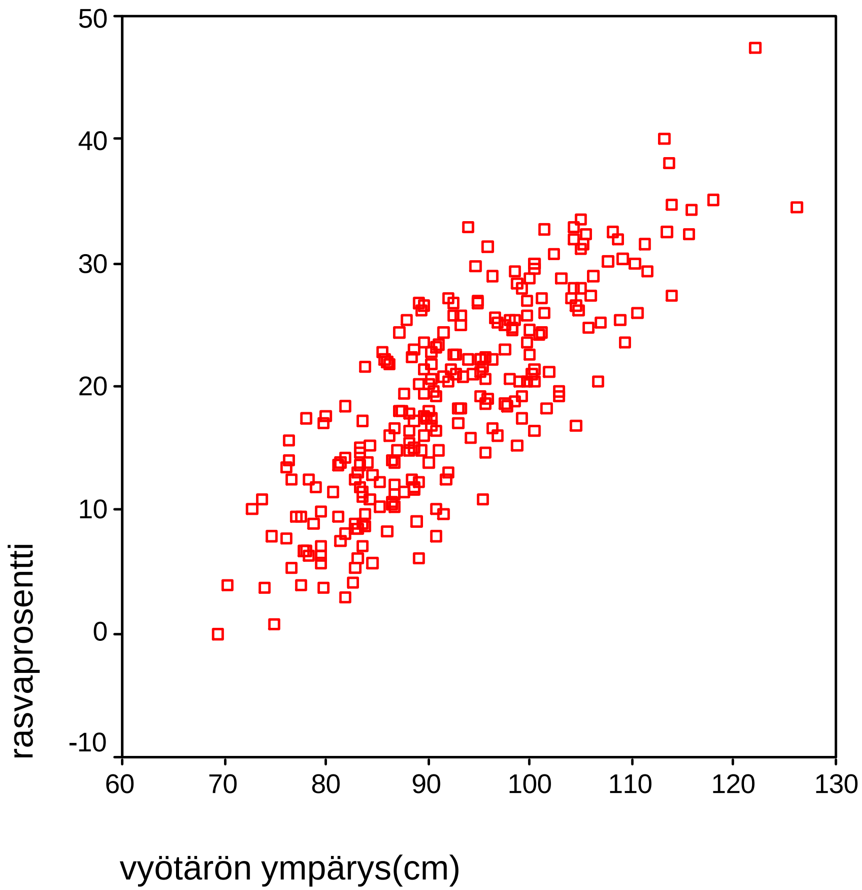
Korrelaatiokerroin r

Mittaa kahden kvantitatiivisen muuttujan välistä lineaarista riippuvuutta, sen voimakkuutta. Mittaa sitä, miten tiiviisti pisteparven pisteet ovat sijoittuneet pisteparveen sovitettavan suoran ympärille.

Ominaisuuksia

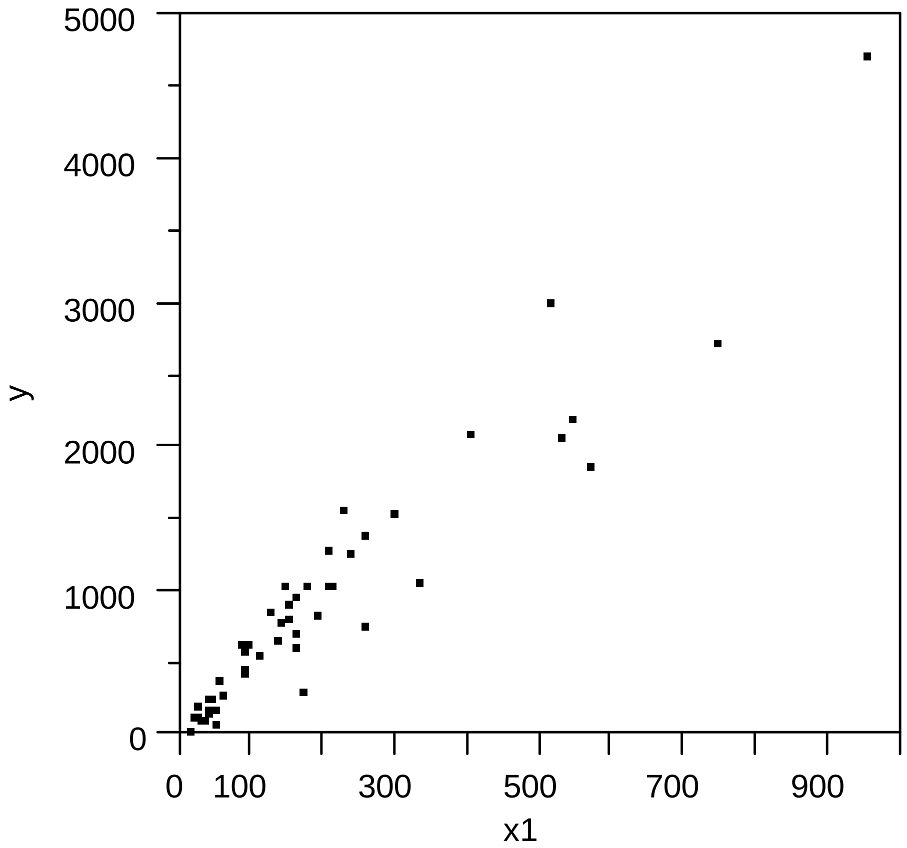
- $-1 \leq r \leq 1$
- $r = 1$, jos kaikki pisteet samalla nousevalla suoralla
- $r = -1$, jos kaikki pisteet samalla laskevalla suoralla
- $r \approx 0$, jos ei lineaarista riippuvuutta

Esim. 5.2.8.



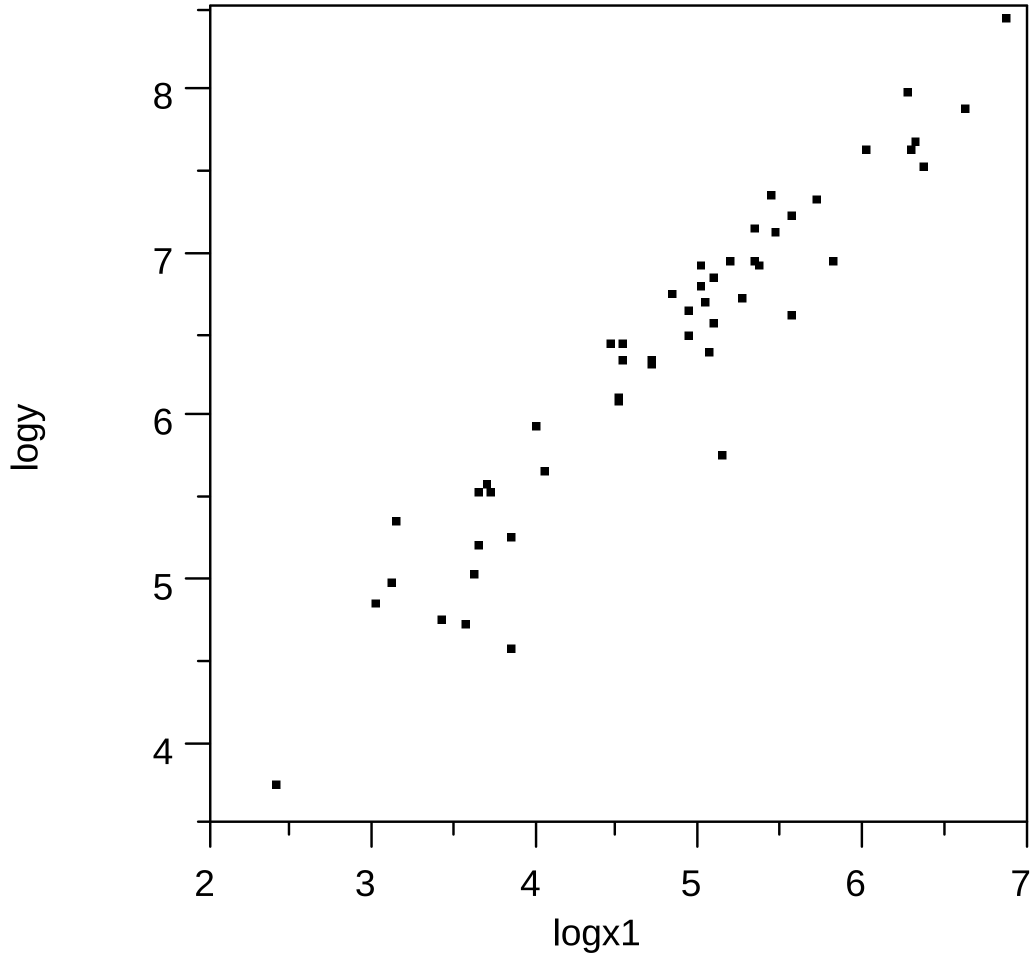
r = 0,825

Esim. 5.2.10.



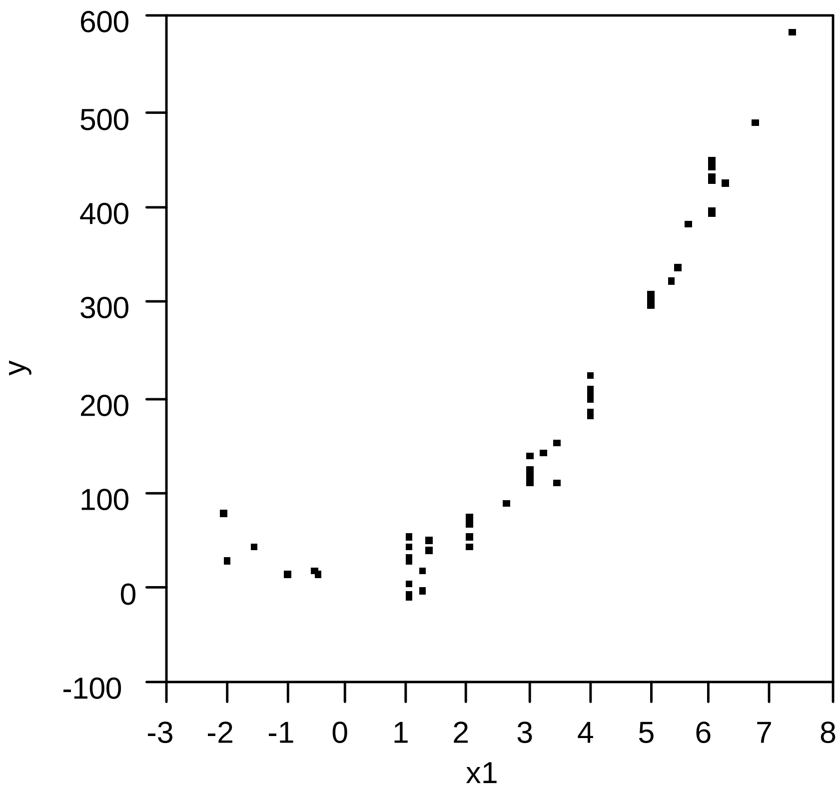
$r = 0{,}9559$

Esim. 5.2.11.

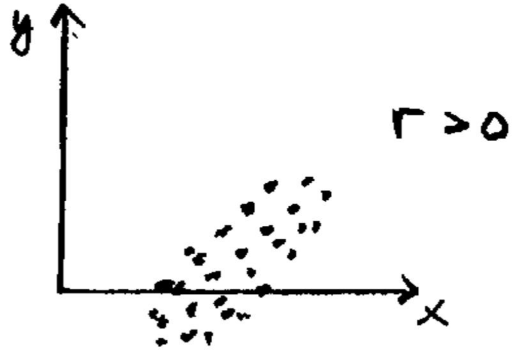


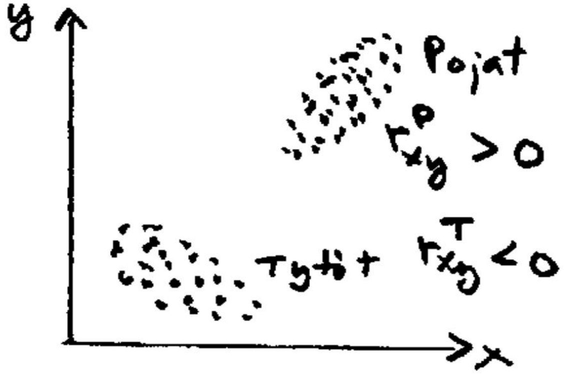
$r = 0{,}9537$

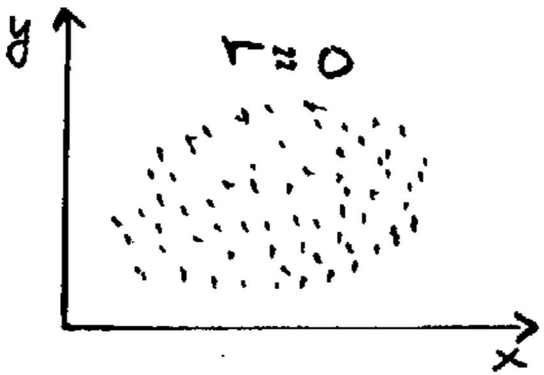
Esim. 5.2.12. Riippuvuutta, joka ei lineaarista.

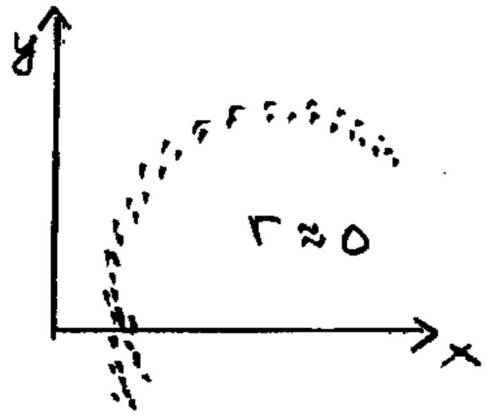


Esim. Pisteparvia ja arviot korrelaatiokertoimista









$r_{xy} > 0$

$|r_{xy}| \leq 1$

Esim. 5.2.13. Pisteparvia ja korrelaatiokertoimia

http://www.sis.uta.fi/tilasto/tiltp7/moniste_8.pdf

Esim. 5.2.17. Korrelaatiomatriisi, CTESTI-aineisto

Correlations

		ika	pituus	paino	cooper
ika	Pearson Correlation	1	,807**	,768**	,399**
	Sig. (2-tailed)		,000	,000	,000
	N	152	152	152	152
pituus	Pearson Correlation	,807**	1	,892**	,236**
	Sig. (2-tailed)	,000		,000	,003
	N	152	153	153	153
paino	Pearson Correlation	,768**	,892**	1	,102
	Sig. (2-tailed)	,000	,000		,210
	N	152	153	153	153
cooper	Pearson Correlation	,399**	,236**	,102	1
	Sig. (2-tailed)	,000	,003	,210	
	N	152	153	153	153

**. Correlation is significant at the 0.01 level (2-tailed).

Korrelaatiokertoimen laskukaava kaavakokoelman kaava (4)

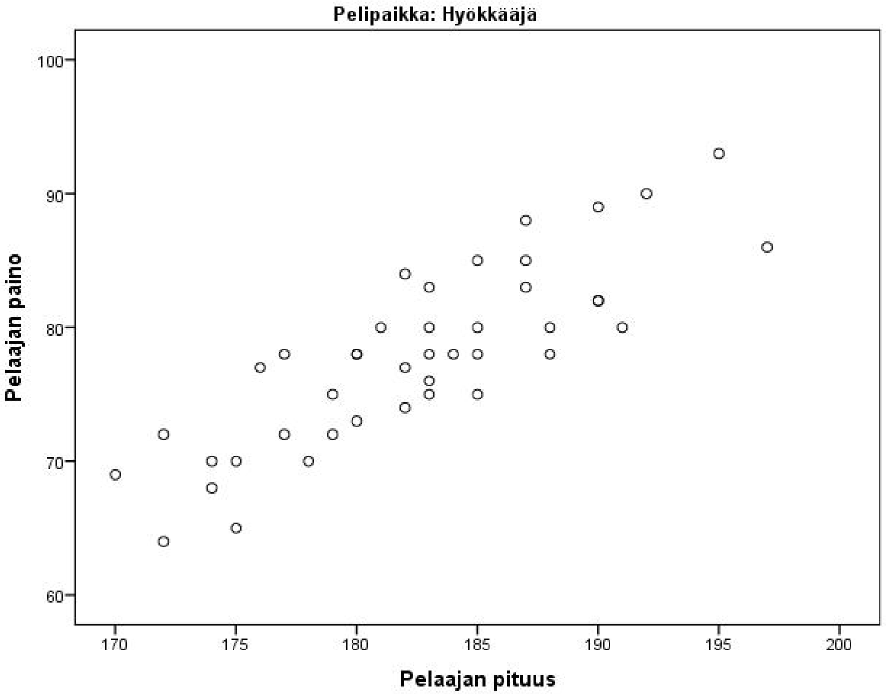
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

ks. myös
http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/esimerkit_kaavoihin.pdf

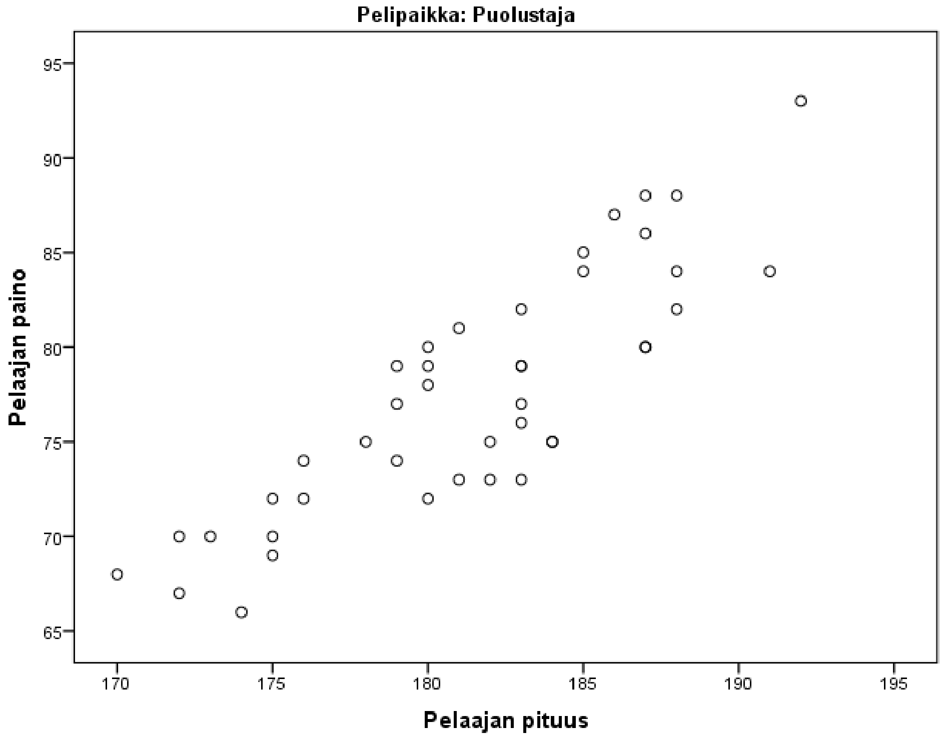
Esim. 5.2.14. Mittayksikön vaihto ei vaikuta korrelaatiokertoimeen, ks. lineaarisen muunnoksen vaikutus korrelaatiokertoimeen

http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=47

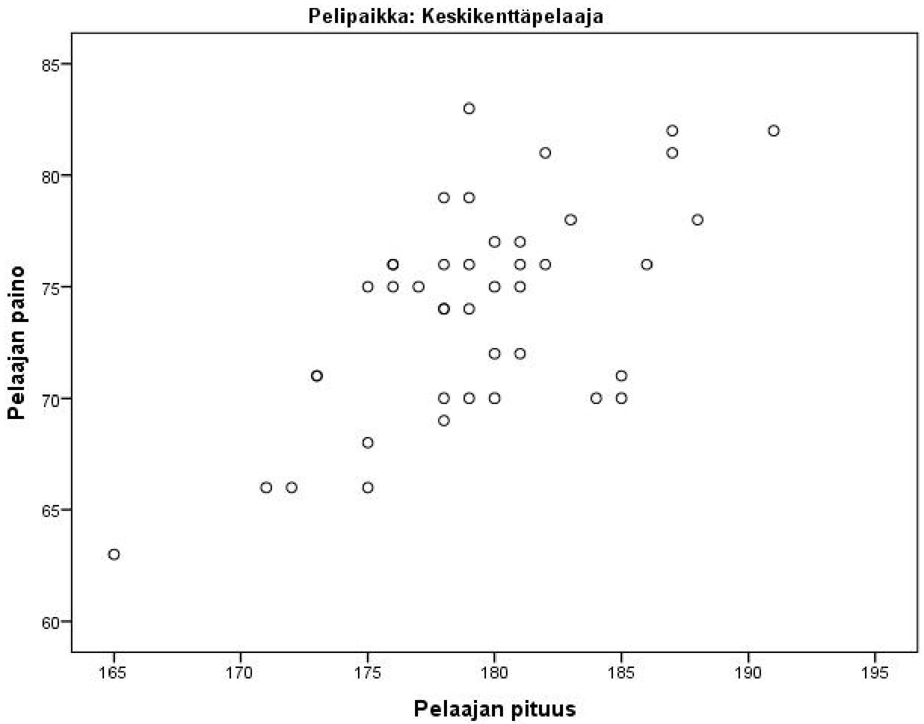
Esim. 5.2.16. Korrelaatiokertoimet pelipaikoittain, ehdolliset korrelaatiot



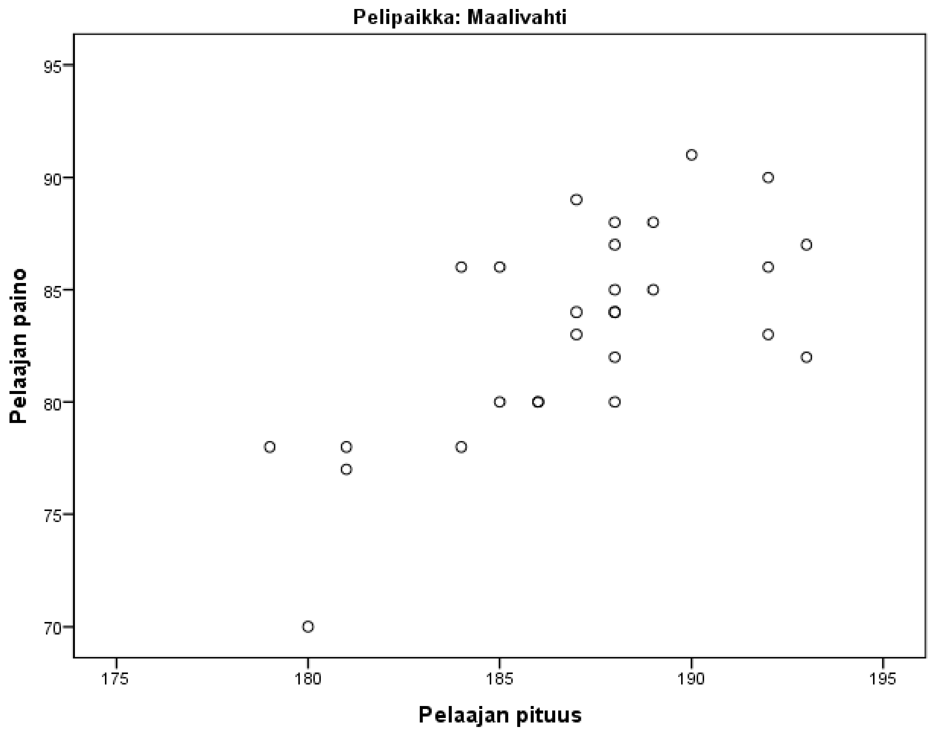
r= 0,84, n=42



r= 0,86, n=42



r= 0,62, n=42



r= 0,68, n=28

Esim. 5.2.17. Osittaiskorrelaatiokertoimet ikä vakioituna, CTESTI-aineisto

Correlations

Control Variables			cooper	paino	pituus
ika	cooper	Correlation	1,000	-,349	-,160
		Significance (2-tailed)	.	,000	,050
		df	0	149	149
	paino	Correlation	-,349	1,000	,719
		Significance (2-tailed)	,000	.	,000
		df	149	0	149
	pituus	Correlation	-,160	,719	1,000
		Significance (2-tailed)	,050	,000	.
		df	149	149	0

6 AIKASARJOISTA

Määritelmä

http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=51

Graafinen esitys

http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=51

Esimerkkejä luentomonisteen esimerkeissä 6.1.1.- 6.1.6.

Harjoitustyön riippuvuustarkastelut

http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/htyop118.pdf#page=4

Riippuvuustarkastelu 1

y (selitettävä) on kvantitatiivinen ja x (selittäjä) kvalitatiivinen

- laatikko-jana-kuvio
- ryhmäkeskiarvot, muut tarvittavat tunnusluvut
- päättely riippumattomien otosten t-testi avulla

Riippuvuustarkastelu 2

y ja x kvalitatiivisia (kvantitatiiviset voi luokitella), selitettävä muuttuja eri kuin riippuvuustarkastelussa 1

- ristiintaulukko
- χ^2–riippumattomuustesti.

7 TILASTOLLISEN PÄÄTTELYN PERUSTEITA

- Miten voidaan arvioida virheellisten komponenttien osuutta tuotannossa?
- Miten voidaan arvioida valmistajan kynttilöiden keskimääräistä palamisaikaa?
- Ovatko kaupungissa eri alueilla myynnissä olevien asuntojen keskineliöhinnat samoja?
- Riippuuko myytävän asunnon kunto sijainnista?
- Miten päättely populaatiosta otoksen perusteella tehdään?

Otos	Populaatio
otoskeskiarvo $\bar{x}$	populaation keskiarvo, odotusarvo μ
otosvarianssi s^2	populaation varianssi σ^2
otoshajonta s	populaation hajonta σ
%-osuus otoksessa p	%-osuus populaatiossa π

Otoksesta määritellyt $\bar{x}$, s^2, s, p ovat otossuureita, joiden käyttäytymistä voidaan arvioida todennäköisyysjakaumien avulla. Näitä jakaumia käytetään hyväksi päättelyssä.

MTTTP1, luento 27.9.2018

7 TILASTOLLISEN PÄÄTTELYN PERUSTEITA

- Miten voidaan arvioida virheellisten komponenttien osuutta tuotannossa?
- Miten voidaan arvioida valmistajan kynttilöiden keskimääräistä palamisaikaa?
- Ovatko kaupungissa eri alueilla myynnissä olevien asuntojen keskineliöhinnat samoja?
- Riippuuko myytävän asunnon kunto sijainnista?
- Miten päättely populaatiosta otoksen perusteella tehdään?

Otos	Populaatio
otoskeskiarvo $\bar{x}$	populaation keskiarvo, odotusarvo μ
otosvarianssi s^2	populaation varianssi σ^2
otoshajonta s	populaation hajonta σ
%-osuus otoksessa p	%-osuus populaatiossa π

Tilastollisessa päättelyssä voidaan arvioida esim.

- odotusarvoa
- prosenttiosuutta
- kahden populaation odotusarvojen yhtäsuuruutta
- muuttujien riippumattomuutta

Otoksesta määritellyt $\bar{x}$, s^2, s, p ovat otossuureita, joiden käyttäytymistä voidaan arvioida todennäköisyysjakaumien avulla. Näitä jakaumia käytetään hyväksi päättelyssä.

7.1 Satunnaisilmiö ja tapahtuma

Esim. 7.1.1. Rahanheitto, nopanheitto, lottoaminen.

Satunnaisilmiö (satunnaiskoe)

useita tulosmahdollisuuksia, epävarmuus tuloksesta

Perusjoukko (E)

kaikki mahdolliset tulokset

Tapahtuma (A)

perusjoukon osajoukko

Esim. 7.1.2.

Rahanheitto

$E = \{kruunu, klaava\}$

tapahtumia

$A = \{kruunu\}$

$B = \{klaava\}$

Nopanheitto

$E = \{1, 2, 3, 4, 5, 6\}$

tapahtuma

$A = \{saadaan\ parillinen\} = \{2,4,6\}$

7.2 Klassinen todennäköisyys

Tapahtuman A todennäköisyys

$$P(A) = k/n$$

n satunnaisilmiön perusjoukon tulosten lukumäärä

k tapahtumaan A liittyvien tulosten lukumäärä

Esim. 7.2.1.

Rahanheitto

$A = \{kruunu\}$

$P(A) = 1/2$

Nopanheitto

$A = \{\text{saadaan parillinen}\} = \{2,4,6\}$

$P(A) = 3/6$

$B = \{1\}$, $P(B) = 1/6$

$D = \{\text{suurempi kuin } 4\} = \{5,6\}$, $P(D) = 2/6$

Tapahtumien A ja B riippumattomuus

7.3 Satunnaismuuttuja ja todennäköisyysjakauma

Esim. 7.3.1. Nopanheitto

X = saatu silmäluku

$P(X=1) = P(X=2) = \ldots = P(X=6) = 1/6$

Esim. 7.3.2. Heitetään kolikkoa neljä kertaa, X = klaavojen lukumäärä heittosarjassa

	lukumäärä		lukumäärä
Kl,Kl,Kl,Kl	4	Kr,Kl,Kl,Kr	2
Kr,Kl,Kl,Kl	3	Kl,Kr,Kl,Kr	2
Kl,Kr,Kl,Kl	3	Kr,Kl,Kr,Kl	2
Kl,Kl,Kr,Kl	3	Kl,Kr,Kr,Kr	1
Kl,Kl,Kl,Kr	3	Kr,Kl,Kr,Kr	1
Kl,Kl,Kr,Kr	2	Kr,Kr,Kl,Kr	1
Kr,Kr,Kl,Kl	2	Kr,Kr,Kr,Kl	1
Kl,Kr,Kr,Kl	2	Kr,Kr,Kr,Kr	0

$P(X=0) = 1/16$, $P(X=3) = 4/16$, $P(X=1) = 4/16$, $P(X=4) = 1/16$, $P(X=2) = 6/16$

Esim. 7.3.4.

Kahden alkion otokset luvuista 1, 2, 3, 4, 5, 6 systemaattisella otannalla ovat {1, 4}, {2, 5}, {3, 6}, joista keskiarvot 2,5, 3,5 ja 4,5, joten

$$P(\bar{X}=2{,}5) = P(\bar{X}=3{,}5) = P(\bar{X}=4{,}5) = 1/3.$$

Satunnaismuuttuja

funktio, joka liittää yksikäsitteisen reaaliluvun jokaiseen tarkasteltavan satunnaisilmiön perusjoukon tulokseen

Diskreetin satunnaismuuttujan X todennäköisyysjakauma

$P(X=x_1) = p_1$, $P(X=x_2) = p_2$...

$p_1 + p_2 + ... = 1$

Jatkuvan satunnaismuuttujan X todennäköisyysjakauma

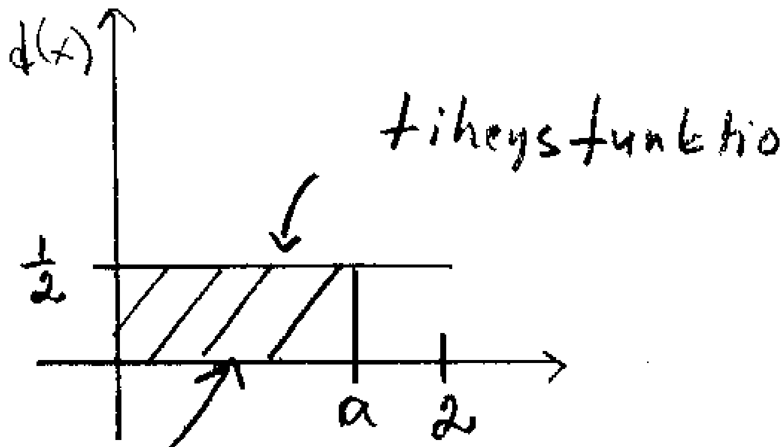
jatkuva funktio $f(x)$, jolle $f(x) \geq 0$ sekä $f(x)$:n ja x-akselin väliin jäävä pinta-ala on yksi.

Funktiota $f(x)$ kutsutaan tiheysfunktioksi.

Satunnaismuuttujan X kertymäfunktio

$F(x) = P(X \leq x)$.

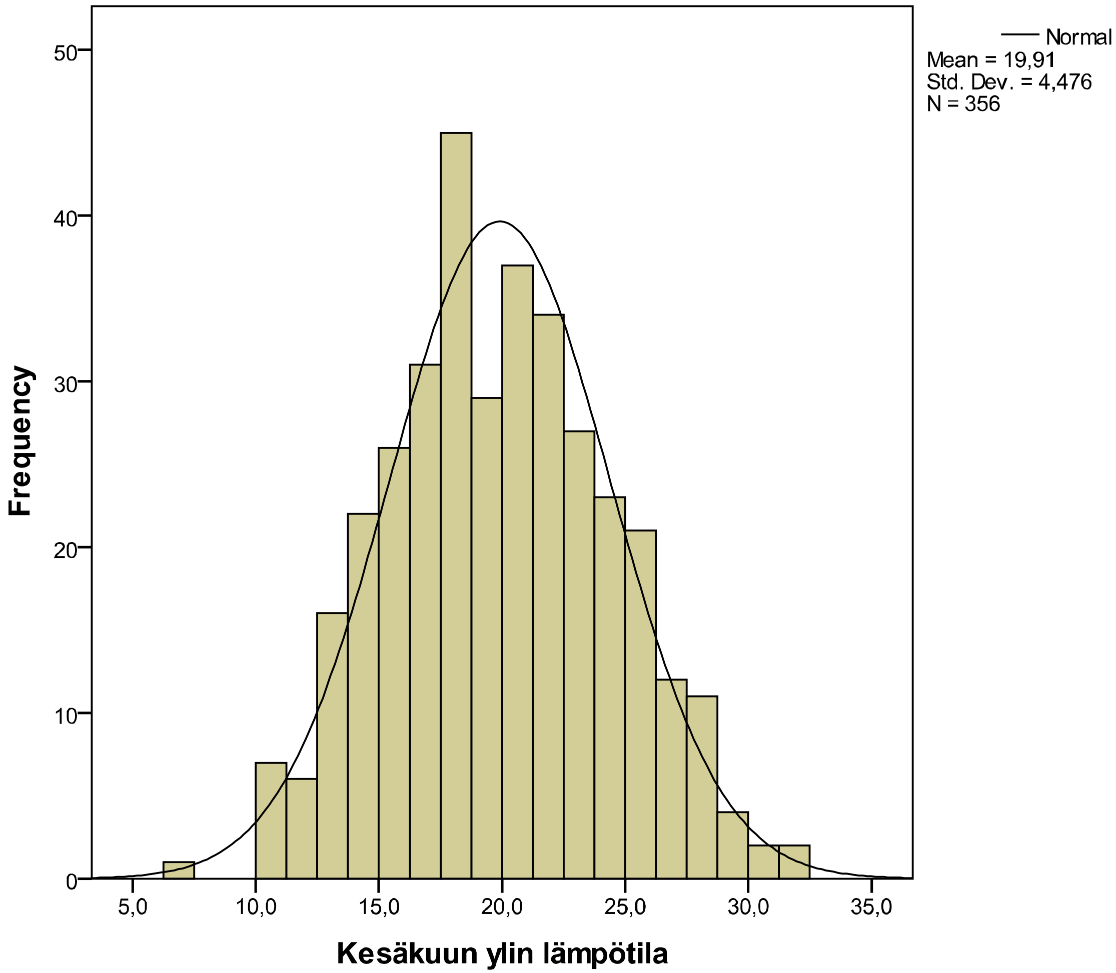
Esim. 7.3.5. Esimerkki erään jatkuvan satunnaismuuttujan tiheys- ja kertymäfunktiosta



$$\text{Pinta-ala} = P(X \le a) = F(a) = a \cdot \frac{1}{2}$$

kertymäfunktio

Esim. Erään tiheysfunktion kuvaaja.



Todennäköisyysjakaumien tunnuslukuja

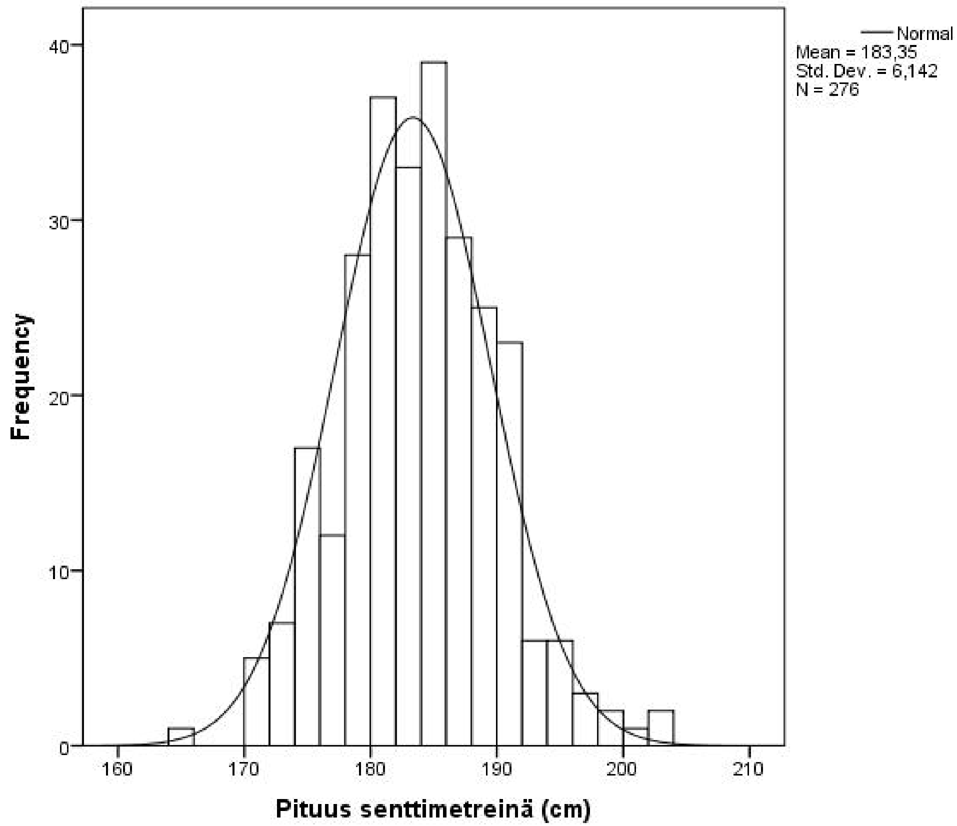
- odotusarvo $E(X) = \mu$
- varianssi $Var(X) = \sigma^2$, keskihajonta σ

Satunnaismuuttujien summat, erotukset, suhteet, jne. ovat myös satunnaismuuttujia.

Satunnaismuuttujien riippumattomuus määritellään vastaavalla tavalla kuin tapahtumien riippumattomuus.

7.4 Normaalijakauma

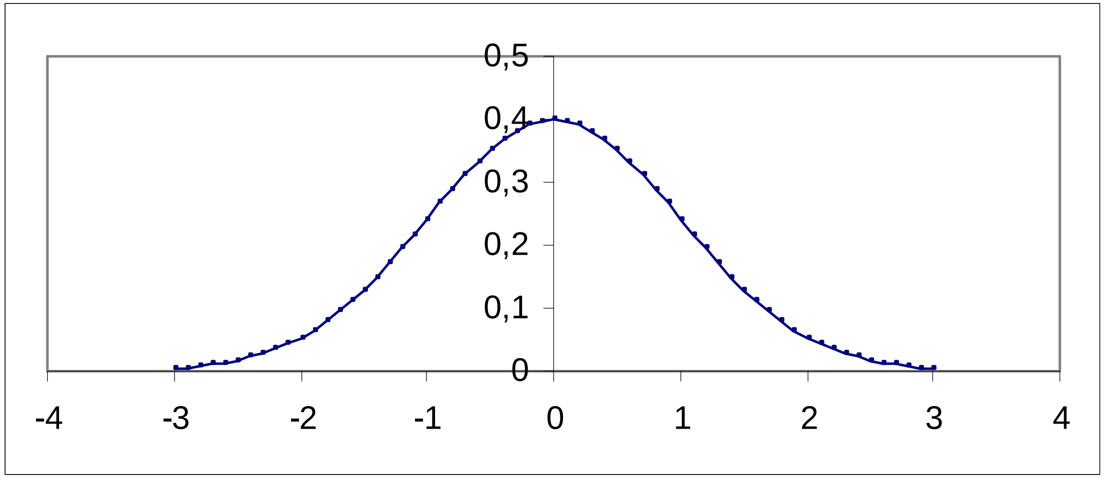
Esim. 7.4.1. Vaahteraliigan pelaajien pituusjakauma. Kuvaan on piirretty normaalijakauman, jonka odotusarvo 183,35 ja varianssi $6{,}142^2$, tiheysfunktio.



Normaalijakauma määritellään parametrein μ ja σ^2, merkitään $X \sim N(\mu, \sigma^2)$, tiheysfunktion kuvaajia, ks. https://fi.wikipedia.org/wiki/Normaalijakauma

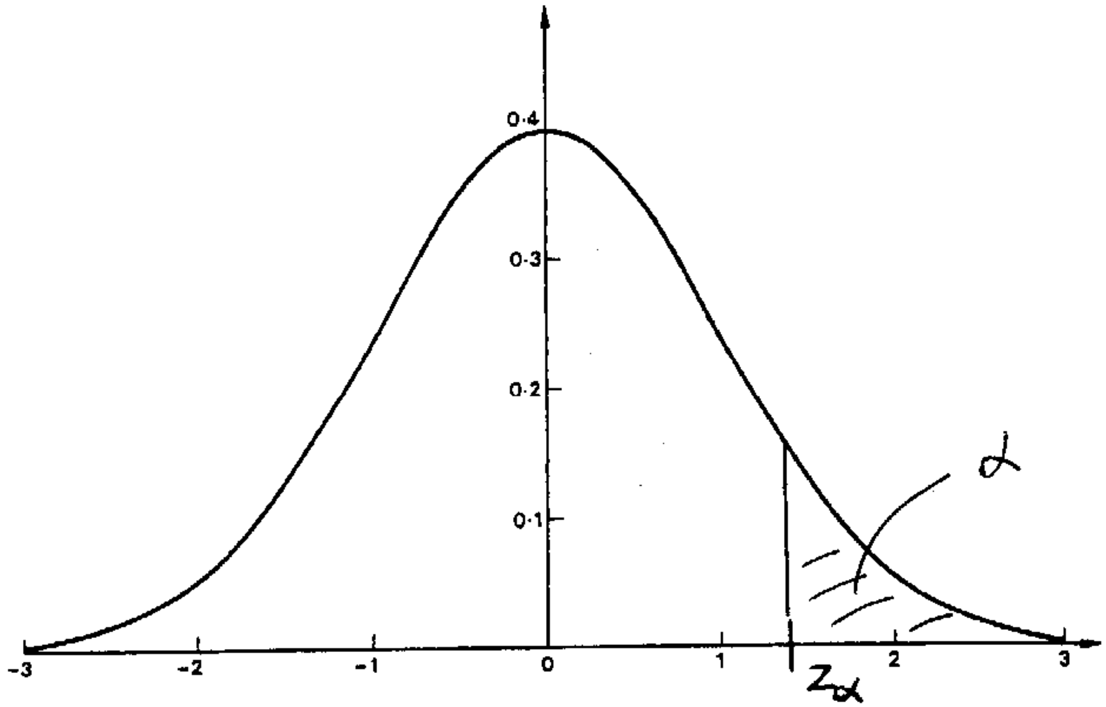
Jos odotusarvo on nolla ja varianssi yksi, kyseessä standardoitu normaalijakauma, merkitään $Z \sim N(0, 1)$. Tällöin $P(Z \leq z) = \Phi(z)$, standardoidun normaalijakauman kertymäfunktiota merkitään $\Phi(z)$:lla.

Esim. 7.4.2. N(0, 1) – jakauman tiheysfunktion kuvaaja

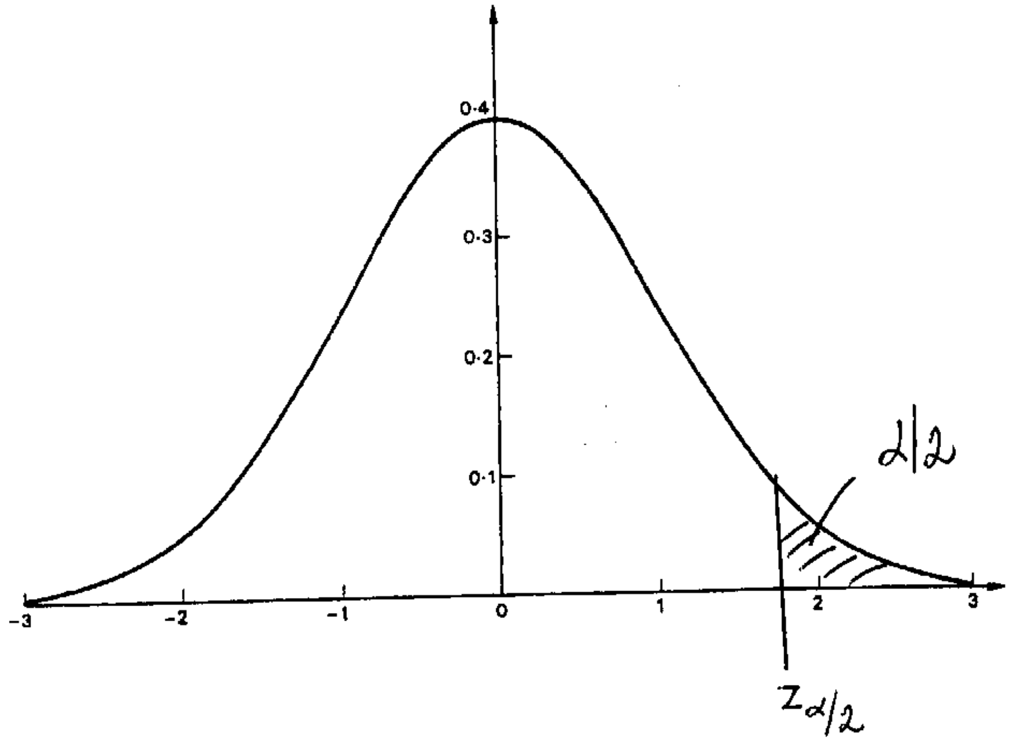


Standardoidun normaalijakauman kertymäfunktion Φ(z) arvoja taulukoitu, ks.
http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/kaavat.pdf

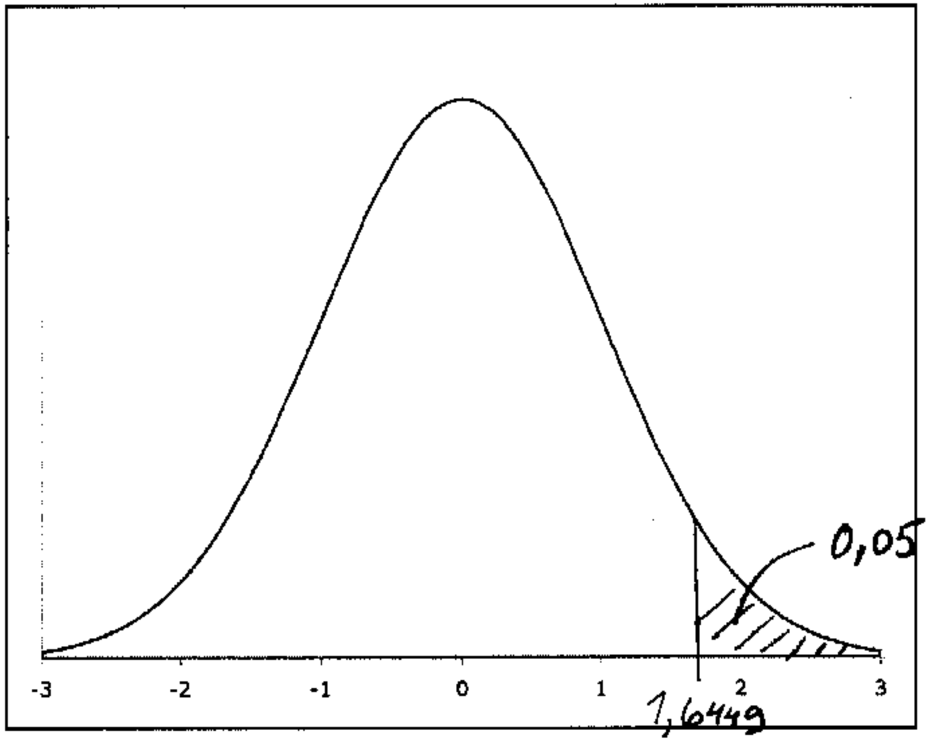
Olkoon Z ~ N(0, 1). Määritellään z_α siten, että $P(Z \geq z_\alpha) = \alpha$, graafisesti:



Samoin $z_{\alpha/2}$ siten, että $P(Z \geq z_{\alpha/2}) = \alpha/2$, graafisesti:

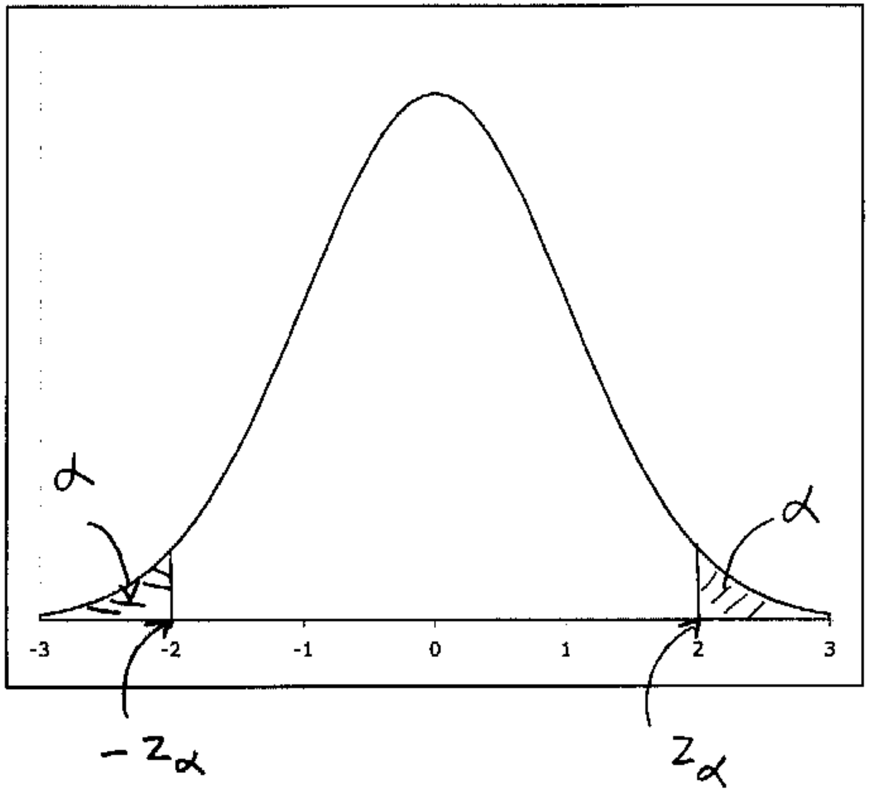


Esim. 7.4.3.



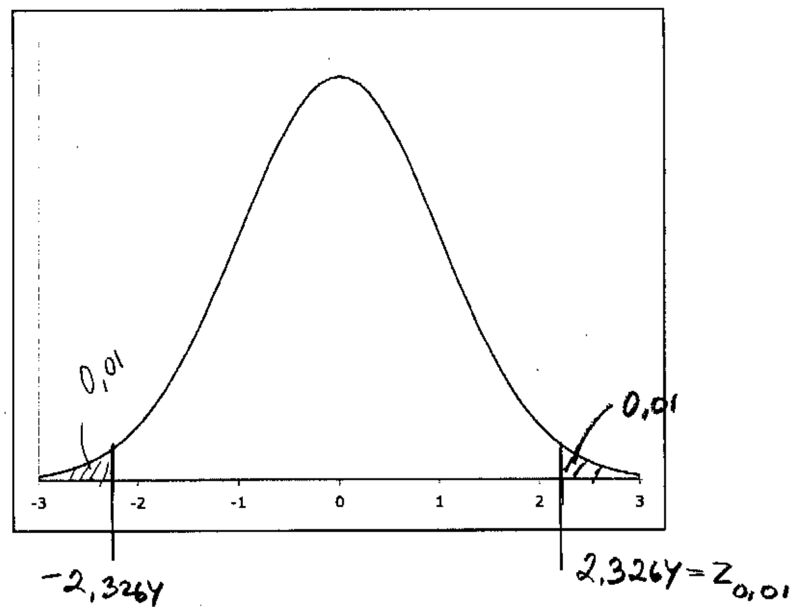
$z_{0,05} = 1,6449$
$z_{0,01} = 2,3264$
$z_{0,05/2} = z_{0,025} = 1,96$

Standardoitu normaalijakauman symmetrinen nollan suhteen



Esim. 7.4.4. Olkoon $Z \sim N(0, 1)$.

$P(Z \geq 2{,}3264) = 0{,}01$, $P(Z \leq -2{,}3264) = 0{,}01$



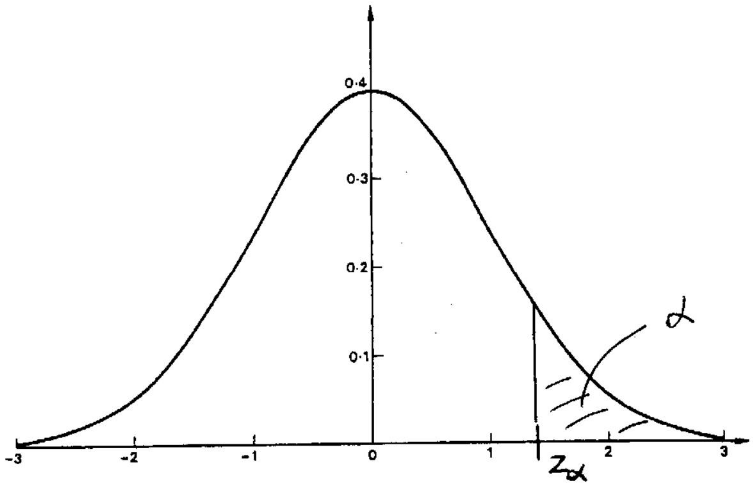
$P(Z \geq 1{,}96) = 0{,}025$, $P(Z \leq -1{,}96) = 0{,}025$

$P(Z \geq 1{,}6449) = 0{,}05$, $P(Z \leq -1{,}6449) = 0{,}05$

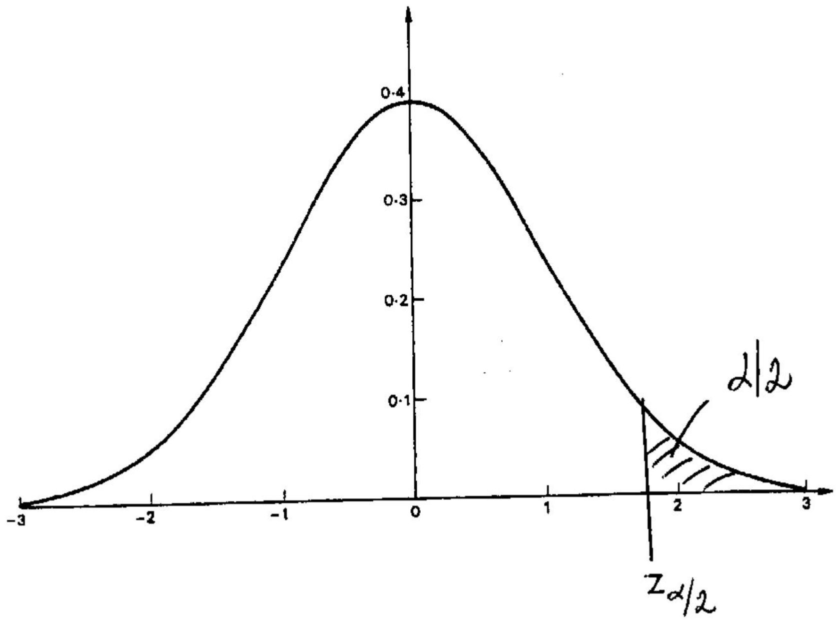
MTTTP1, luento 2.10.2018

7.4 Normaalijakauma (kertausta)

Olkoon $Z \sim N(0, 1)$. Määritellään z_α siten, että $P(Z \geq z_\alpha) = \alpha$, graafisesti:



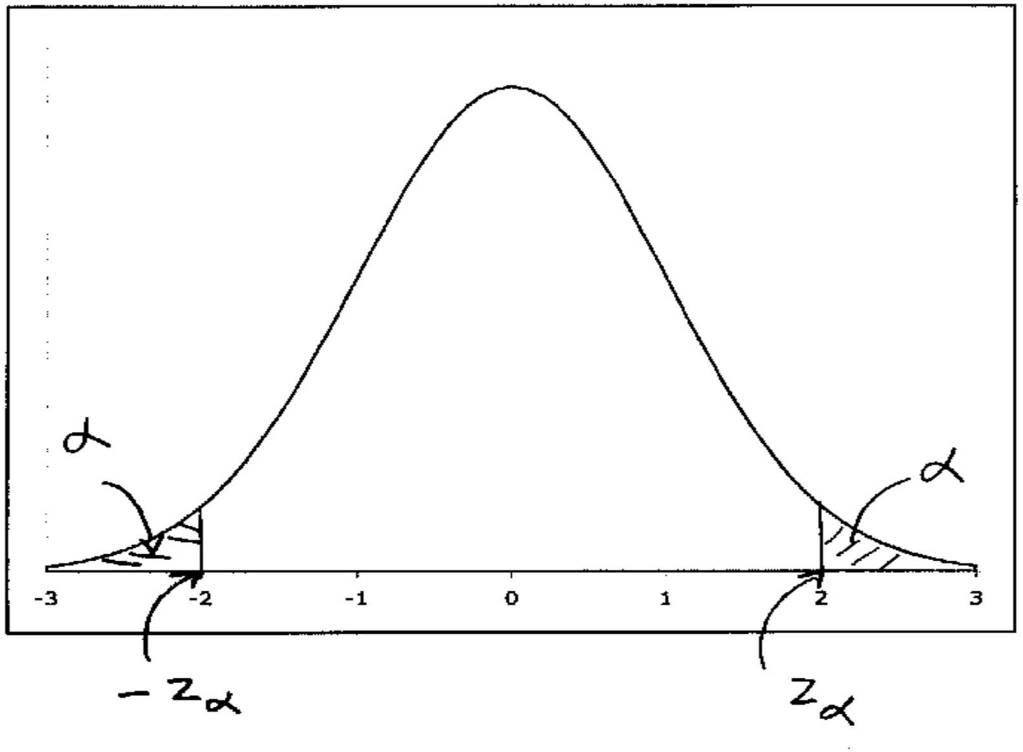
Samoin $z_{\alpha/2}$ siten, että $P(Z \geq z_{\alpha/2}) = \alpha/2$, graafisesti:



Esim.

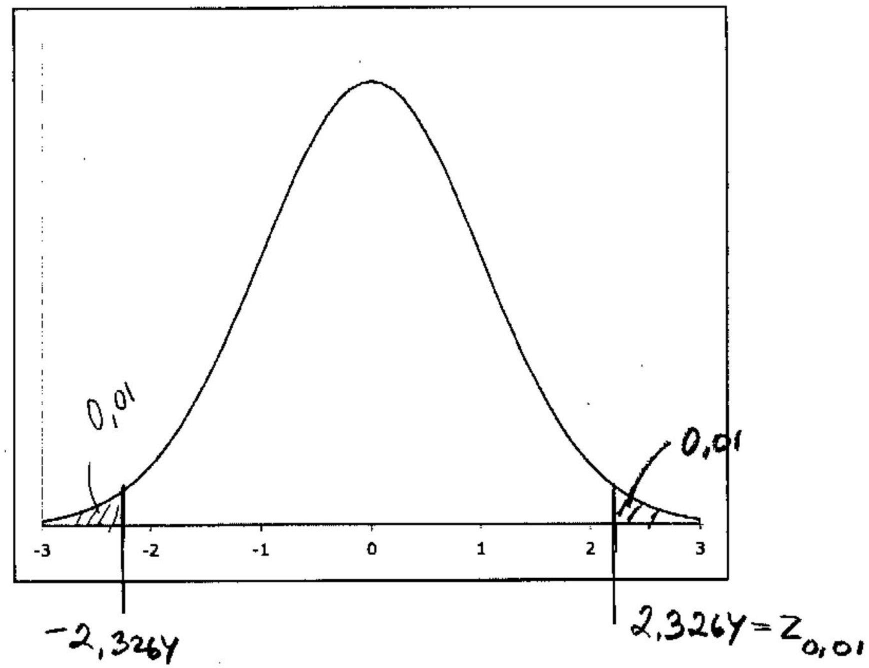
$$z_{0,05} = 1{,}6449$$
$$z_{0,01} = 2{,}3264$$
$$z_{0,05/2} = z_{0,025} = 1{,}96$$

Standardoitu normaalijakauman symmetrinen nollan suhteen



Esim. 7.4.4. Olkoon $Z \sim N(0, 1)$.

$P(Z \geq 2{,}3264) = 0{,}01$, $P(Z \leq -2{,}3264) = 0{,}01$



$P(Z \geq 1{,}96) = 0{,}025$, $P(Z \leq -1{,}96) = 0{,}025$

$P(Z \geq 1{,}6449) = 0{,}05$, $P(Z \leq -1{,}6449) = 0{,}05$

7.5 Satunnaisotos, otossuure ja otantajakauma

Päätelmät populaatiosta otoksen perusteella

- puolueen kannatus
- kynttilöiden keskimääräinen palamisaika
- asuntojen keskimääräiset neliöhinnat keskustassa ja lähiössä

Miten päättely tehdään? Miten tulosten luotettavuutta voidaan arvioida?

Päättely tehdään satunnaisotoksen perusteella.

Satunnaismuuttujajono $X_1, X_2, ..., X_n$ on satunnaisotos, jos X_i:t ovat riippumattomia ja noudattavat samaa jakaumaa.

Esim. Satunnaisotos $X_1, X_2, ..., X_n$ normaalijakaumasta $N(\mu, \sigma^2)$. Tällöin jokainen X_i noudattaa normaalijakaumaa parametrien μ, σ^2 ja X_i:t ovat toisistaan riippumattomia.

Otossuure on satunnaisotoksen perusteella määritelty funktio.

Olkoon satunnaisotos X_1, X_2, ..., X_n normaalijakaumasta $N(\mu, \sigma^2)$, tällöin

$$\bar{X} \sim N(\mu, \sigma^2/n), \text{ kaava (6).}$$

Otoskeskiarvo on otossuure, jonka todennäköisyysjakauma tiedetään. Se on normaalijakauma, havainnollistaminen simuloiden

http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Olkoon populaatiossa π % tietyn tyyppisiä alkioita ja p = tietyn tyyppisten alkioiden % -osuus otoksessa.

Tällöin

$$p \sim N(\pi, \pi(100-\pi)/n), \text{ likimain, kaava (7).}$$

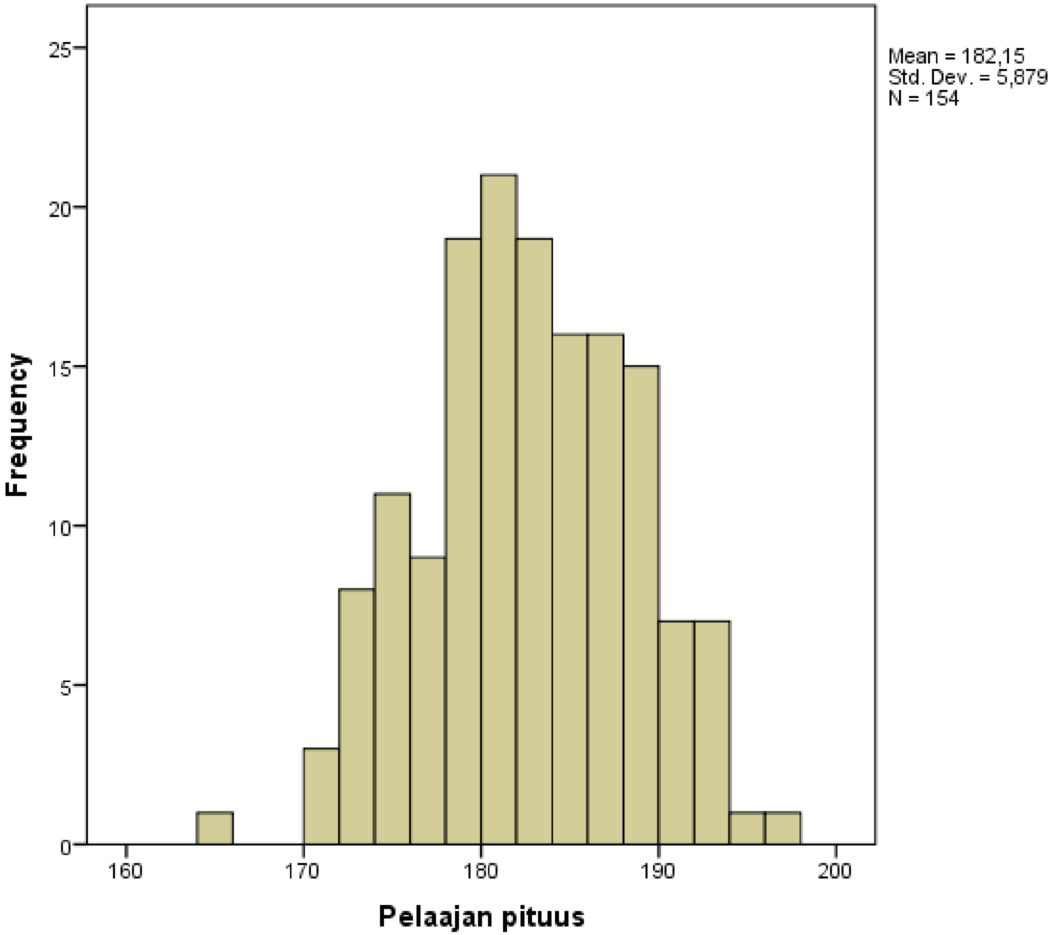
Viallisten prosenttiosuus otoksessa (p) on otossuure, jonka jakauma on likimain normaalijakauma.

Otossuureiden jakaumia käytetään päättelyyn liittyvien tulosten luotettavuuden arvioinnissa.

7.6 Piste-estimointi ja luottamusvälejä

Esim. Vuonna 2007 suomalaisen miesten keskipituuden arvioitiin olevan 179,6 cm, naisten 165,9, http://fi.wikipedia.org/wiki/Ihmisen_pituus#Ihmisten_keskipituus_eri_maissa

Esim. Jalkapalloilijat 2006, jalkapalloilijoiden keskipituuden arviointi. Arvioidaan keskipituuden olevan 182,15 cm.



Esim. Puolueen kannatusarviot,
https://yle.fi/uutiset/3-10387592 (6.9.2018)

Arvioidaan SDP:n kannatuksen olevan 20,3 %.

Estimointi

populaation tuntemattoman parametrin arviointia otossuureen avulla (piste-estimointi)

Estimaattori

otossuure, jolla estimoidaan tuntematonta parametria

Estimaatti

estimaattorin arvo (tehdyn otoksen perusteella laskettu)

Estimaattorin keskivirhe

estimaattorin hajonta

Estimoitava parametri	Esti-maattori	Estimaattorin keskivirhe	Estimoitu keskivirhe
μ	$\bar{X}$	$\sigma/\sqrt{n}$	$s/\sqrt{n}$
π	p	$\sqrt{\pi(100-\pi)/n}$	$\sqrt{p(100-p)/n}$
σ	s		

Esim. Puolueen kannatuksen arviointi

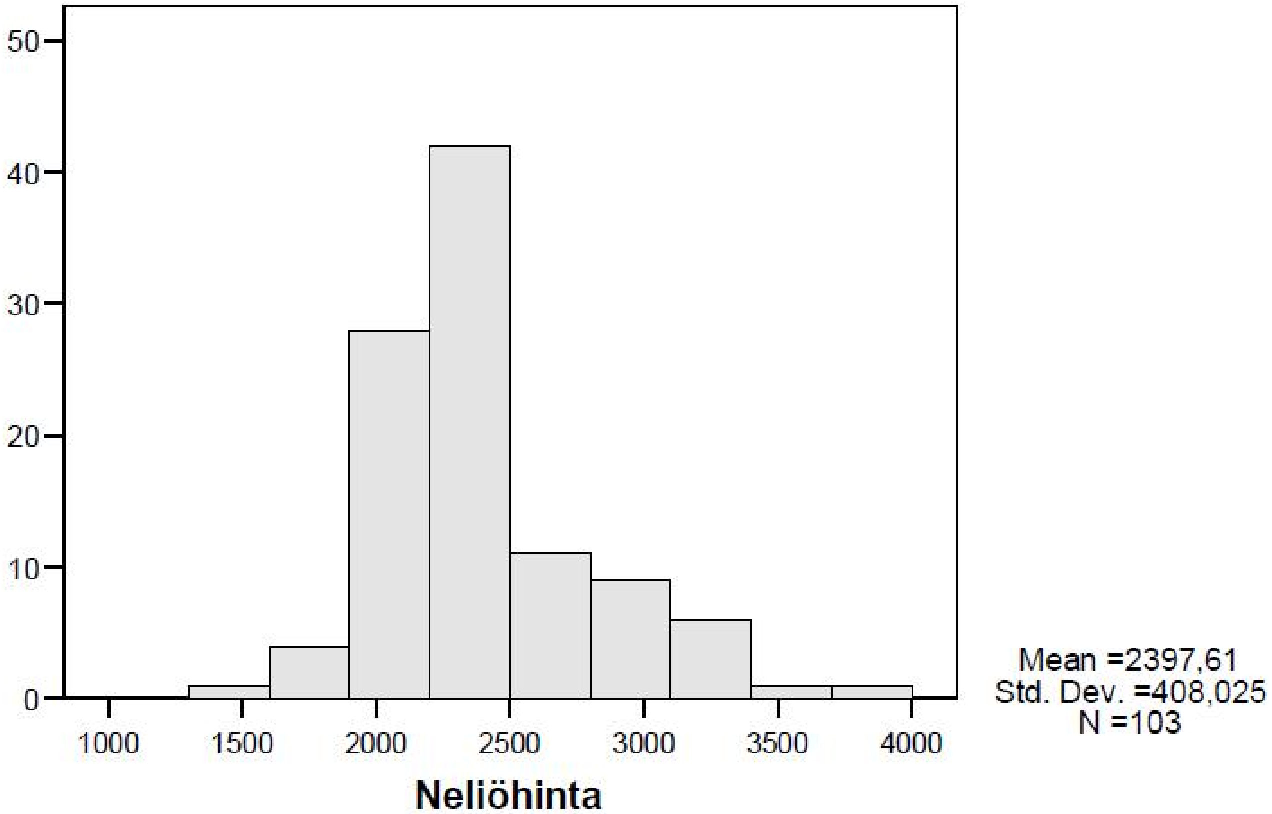
$p = 20{,}3$ %,

$n = 1460$ (kantansa ilmoittaneet).

Kannatuksen estimoitu keskivirhe

$$\sqrt{20{,}3(100-20{,}3)/1460} = 1{,}05.$$

Esim. 7.6.1. Kerrostalohuoneistojen keskimääräisen neliöhinnan estimointi, $\bar{x}$ = 2398, s = 408, joten estimoitu keskivirhe on $408/\sqrt{103}$ = 40,2.



Esim. Jalkapalloilijoiden keskipituuden estimoitu keskivirhe $5{,}879/\sqrt{154} = 0{,}474$.

Myös nk. luottamisvälin avulla voidaan arvioida populaation tuntematonta parametria, tällöin kyse väliestimoinnista.

Muodostetaan väli, joka peittää parametrin etukäteen valitulla todennäköisyydellä, nk. luottamustasolla.

Luottamusväli on satunnaisväli, joka sisältää estimoitavan parametrin todennäköisyydellä $1 - \alpha$. Valitaan α esim. 0,05 tai 0,01. Tällöin kyse 95 %:n tai 99 %:n luottamusvälistä.

7.6.1 Prosenttiosuuden luottamusväli

Kaava (8), 100(1 - α) %:n luottamusväli prosenttiosuudelle

$$p \pm z_{\alpha/2}\sqrt{p(100-p)/n}$$

95 %:n luottamusväli, α = 0,05, $z_{0,05/2} = z_{0,025} = 1,96$

99 %:n luottamusväli, α = 0,01, $z_{0,01/2} = z_{0,005} = 2,5758$

Esim. 7.6.4. Satunnaisesti valituista 100 henkilöstä puoluetta kannatti 18 %. Puolueen kannatuksen 95 %:n luottamusväli

$$18 \pm 1{,}96\sqrt{18(100-18)/100}$$

$$18 \pm 7{,}5$$

Arvioidaan kannatuksen olevan välillä 10,5 – 25,5. Virhemarginaali ±7,5 %-yksikköä.

Esim. Puolueen kannatusarviot ja virhemarginaali, Puolueen kannatusarviot, Puolueen kannatusarviot, https://yle.fi/uutiset/3-10387592 (6.9.2018)

Esim. Kahvin myyjä väittää, että 15 % kahvin juojista valitsee kahvimerkin hinnan perusteella. Tutkitaan myyjän väitettä. Tehdään tutkimus, jossa 250 kahvin juojalta kysytään kahvimerkin valintaan vaikuttavia tekijöitä. Vastanneista 25 valitsi kahvinsa hinnan perusteella. Uskotko myyjän väitteen?

Nyt $n = 250$, $p = 100 \cdot 25/250 = 10$

95 %:n luottamusväli hinnan perusteella valintansa tekevien prosenttiosuudelle

$$10 \pm 1{,}96\sqrt{10(100-10)/250}$$

$$10 \pm 3{,}7$$

Koska 15 ei kuulu luottamusvälille, ei uskota väitettä.

99 %:n luottamusväli

$$10 \pm 2{,}5758\sqrt{10(100-10)/250}$$

$$10 \pm 4{,}9,$$

sama päättely.

Sivut 20-24 seuraavalle luennolle

7.6.2 Populaation odotusarvon luottamusväli

Esim. 7.6.6. Arvioidaan poikien keskimääräistä syntymäpituutta, siis poikapopulaation keskiarvoa. Otoksessa 65 pojan syntymäpituuden keskiarvo 50,95 cm ja keskihajonta 1,97 cm. Arvio populaation odotusarvon luottamusvälin avulla, määrittämisessä käytetään otoskeskiarvoa ja otoshajontaa. Poikien keskipituuden arvellaan olevan välillä 50,5 cm – 51,4 cm.

SPSS-tulos:

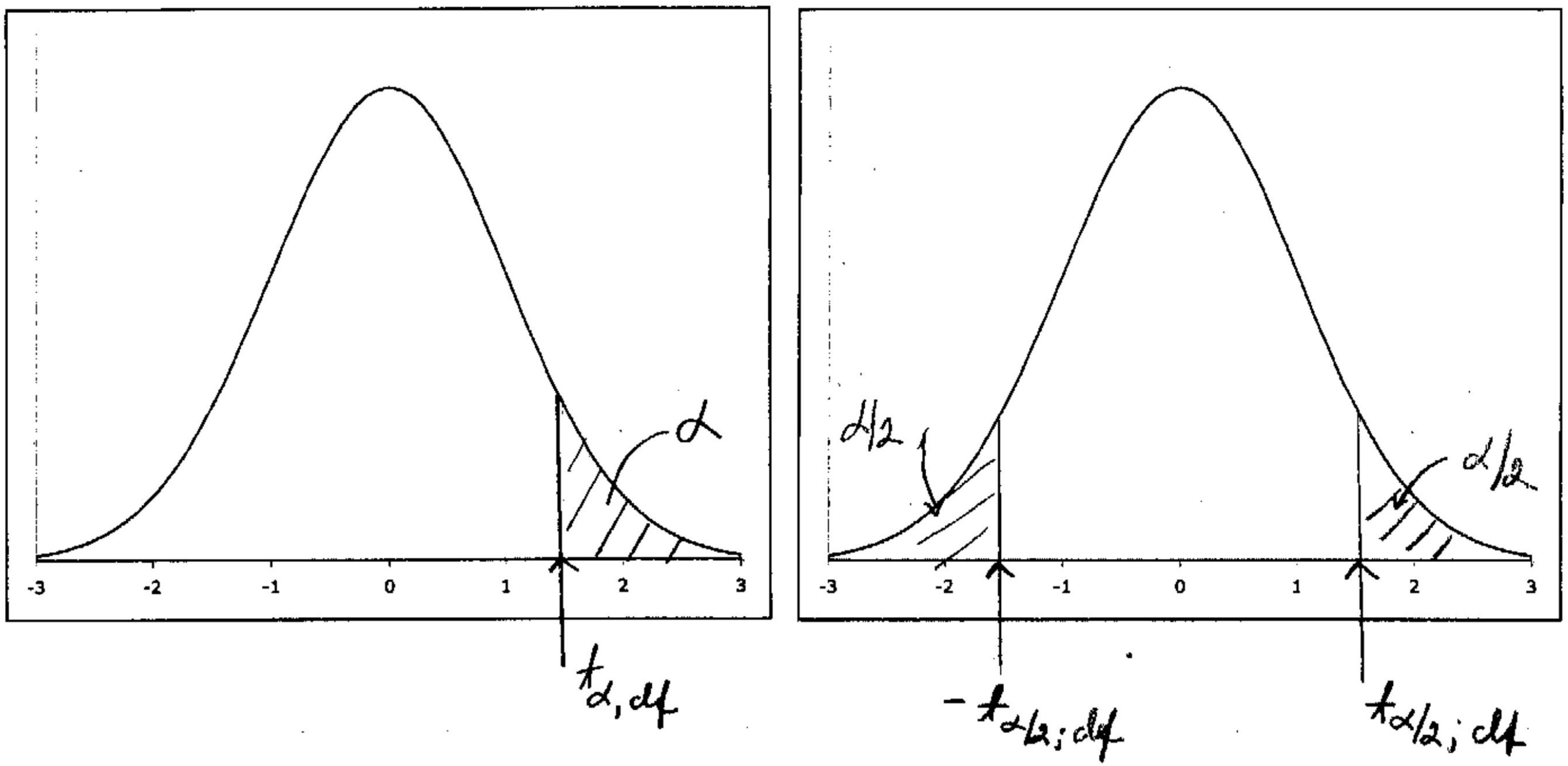
Descriptives

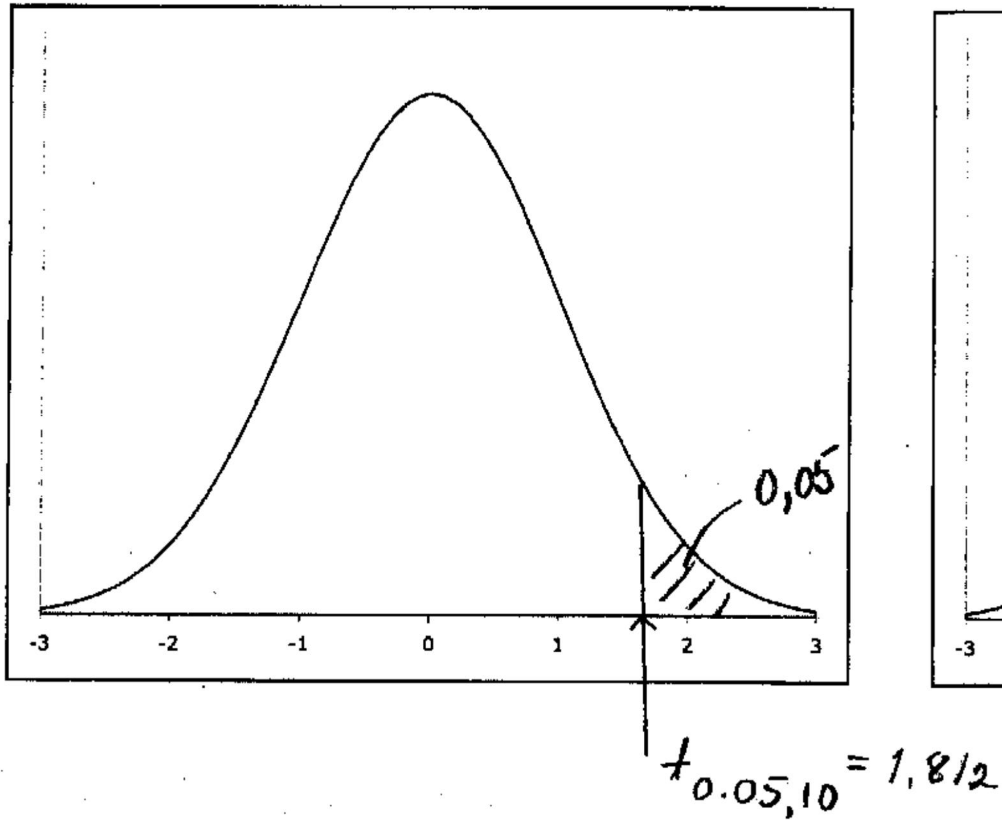
			Statistic	Std. Error
pituus	Mean		50,95	,245
	95% Confidence Interval for Mean	Lower Bound	50,47	
		Upper Bound	51,44	
	Std. Deviation		1,972	

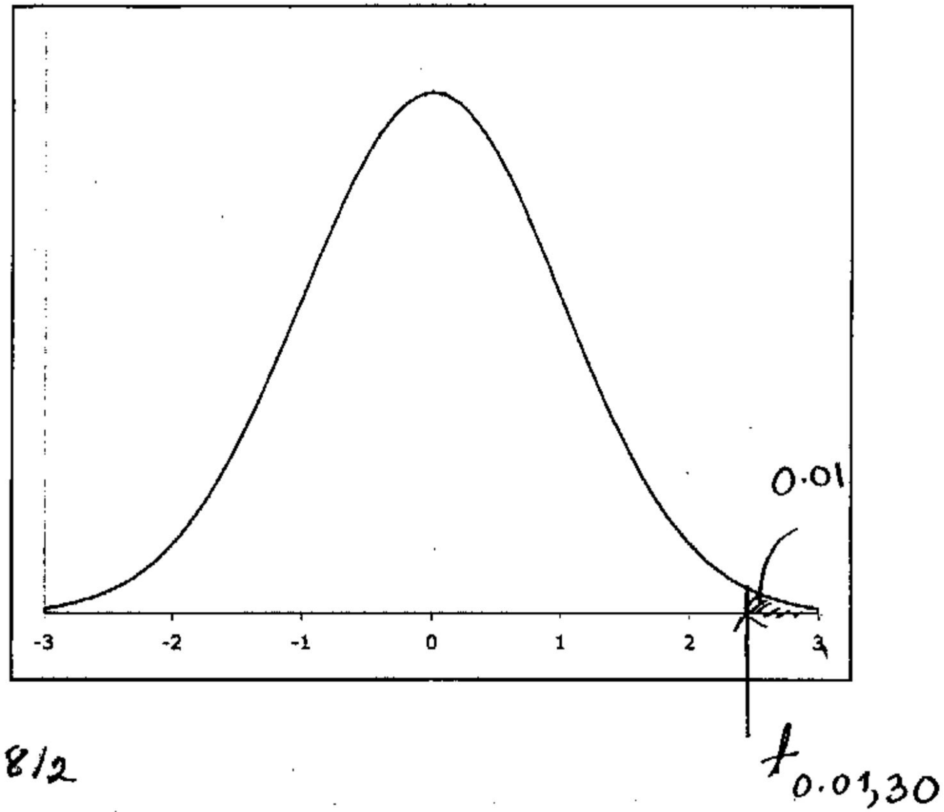
Kaava (9), 100(1 - α) %:n luottamusväli odotusarvolle

$$\overline{X} \pm t_{\alpha/2;n-1} s/\sqrt{n}$$

Studentin t-jakauman taulukkoarvot $t_{\alpha,df}$ ja $t_{\alpha/2,df}$




0,05
-3
-2
-1
0
1
2
3
t0.05,10 = 1,812


0.01
-3
-2
-1
0
1
2
3
t0.01,30

Esim. 7.6.9. Tiedetään, että eräs kirjailija käyttää tuotannossaan virkkeitä, joiden keskipituus on 32 sanaa. Tutkija lukee erään tekstin, jossa on 30 virkettä. Näiden 30 virkkeen keskipituus on 35,5 sanaa ja keskihajonta 6,8 sanaa. Voisiko teksti olla peräisin kyseisen kirjailijan tuotannosta?

Muodostetaan odotusarvon 95 %:n luottamusväli. Nyt $t_{0,05/2;30-1} = 2{,}045$ ja luottamusväli $35{,}0 \pm 2{,}045 \cdot 6{,}8/\sqrt{30}$. Saadaan väliksi 32,5 – 37,5, jolle 32 ei kuulu. Päätellään, että teksti ei ole kyseisen kirjailijan tuotantoa.

MTTTP1, luento 4.10.2018

KERTAUSTA

Luottamisvälin avulla voidaan arvioida populaation tuntematonta parametria.

Muodostetaan väli, joka peittää parametrin etukäteen valitulla todennäköisyydellä, nk. luottamustasolla.

Olkoon populaatiossa π % ”viallisia”.

Nyt π:n 100(1 - α) %:n luottamusväli on

$$p \pm z_{\alpha/2}\sqrt{p(100-p)/n}$$

kaava (8)

Esim. Rahapelin pitäisi antaa voitto 20 %:lle pelatuista peleistä. Pelaat peliä 200 kertaa ja voitat 32 kertaa. Voitko uskoa, että 20 % peleistä voittaa?

Muodostetaan 95 %:n luottamusväli prosenttiosuudelle. Nyt α = 0,05, $z_{0,05/2} = z_{0,025}$ = 1,96, n = 200, p = 16, luottamusväli

$$16 \pm 1{,}96\sqrt{16(100-16)/200}$$

$$16 \pm 5{,}1$$

Koska 20 kuuluu luottamusvälille, päätellään pelin toimivan luvatulla tavalla.

Esim. 7.6.5. Yritys valvoo tuotantoaan. Virheellisten komponenttien osuus ei saisi olla suurempi kuin 4 %. Laaduntarkkailussa tehtiin 500 komponentin otos, jossa 28 komponenttia osoittautui virheellisiksi. Onko tuotanto keskeytettävä?

95 %:n luottamusväli virheellisten komponenttien prosenttiosuudelle

$$5{,}6 \pm 1{,}96\sqrt{5{,}6(100-5{,}6)/500}$$

Virheellisten osuuden arvellaan olevan välillä 3,6 % - 7,6 %, joten vaihtelu on sallituissa rajoissa, koska 4 % kuuluu luottamusvälille.

7.6.2 Populaation odotusarvon luottamusväli

Esim. 7.6.6. Arvioidaan poikien keskimääräistä syntymäpituutta, siis poikapopulaation keskiarvoa. Otoksessa 65 pojan syntymäpituuden keskiarvo 50,95 cm ja keskihajonta 1,97 cm. Arvio populaation odotusarvon luottamusvälin avulla, määrittämisessä käytetään otoskeskiarvoa ja otoshajontaa. Poikien keskipituuden arvellaan olevan välillä 50,5 cm – 51,4 cm.

SPSS-tulos:

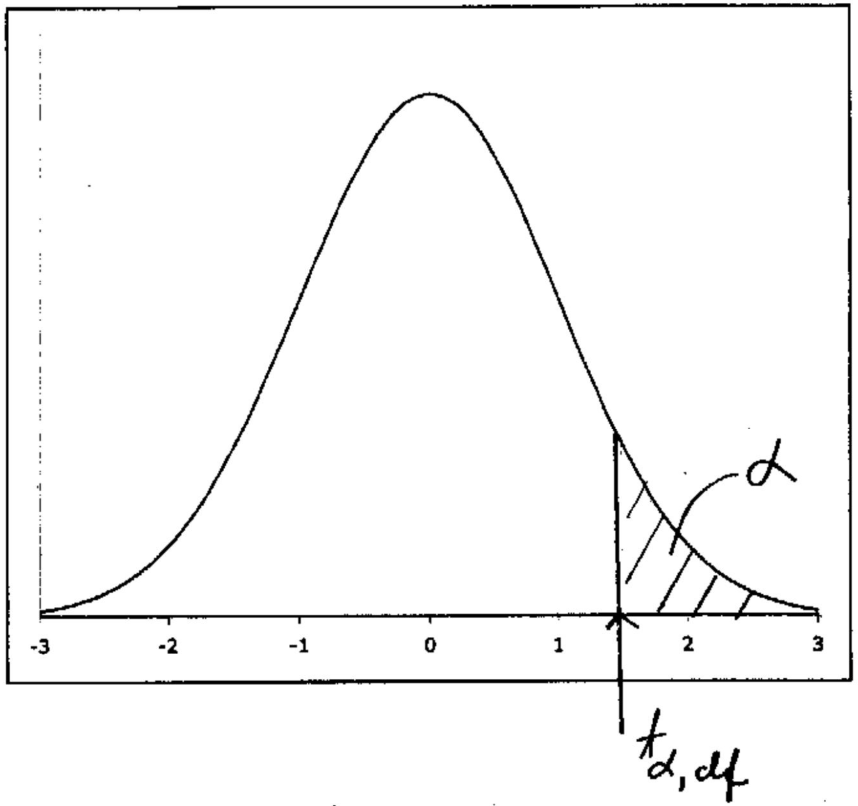
Descriptives

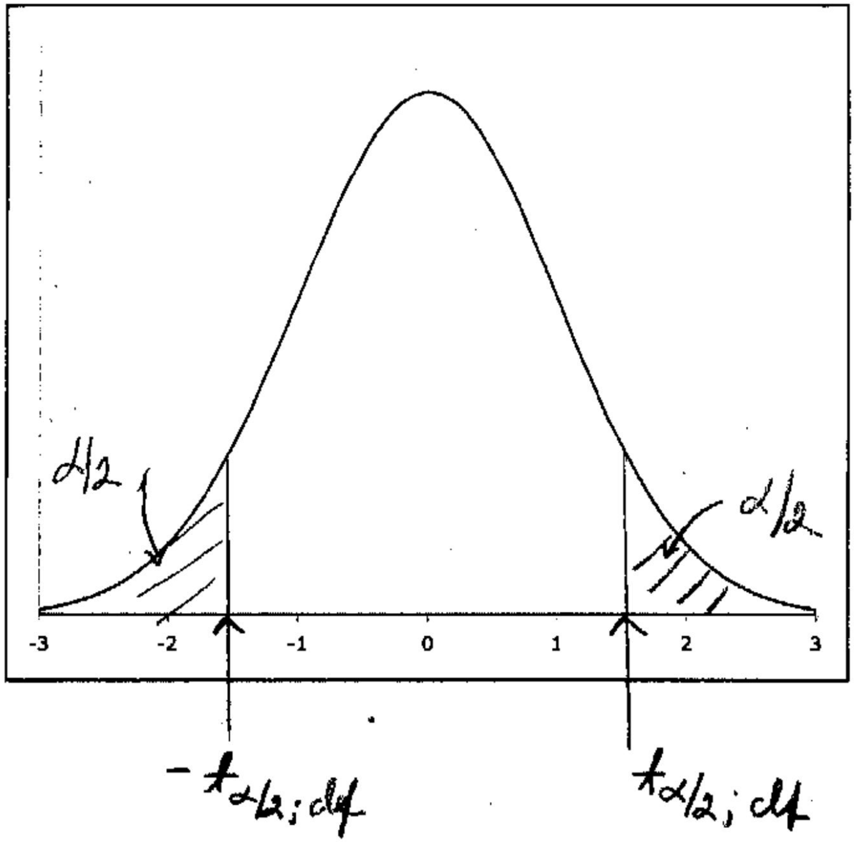
			Statistic	Std. Error
pituus	Mean		50,95	,245
	95% Confidence Interval for Mean	Lower Bound	50,47	
		Upper Bound	51,44	
	Std. Deviation		1,972	

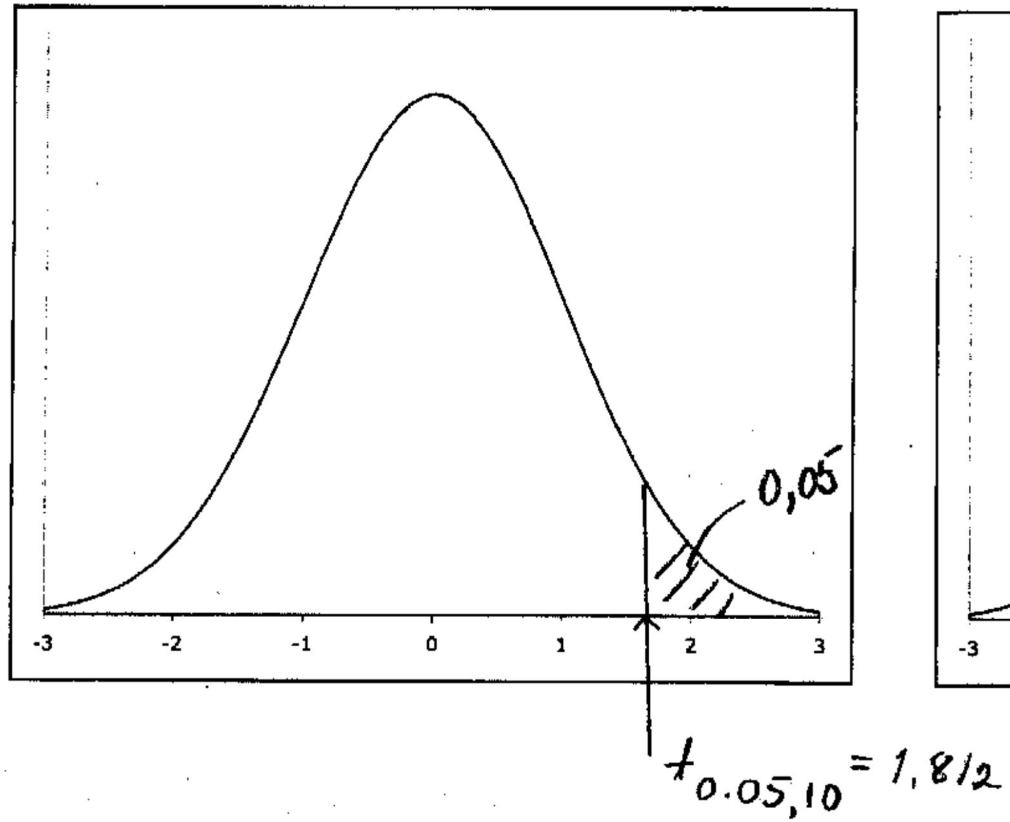
Kaava (9), 100(1 - α) %:n luottamusväli odotusarvolle

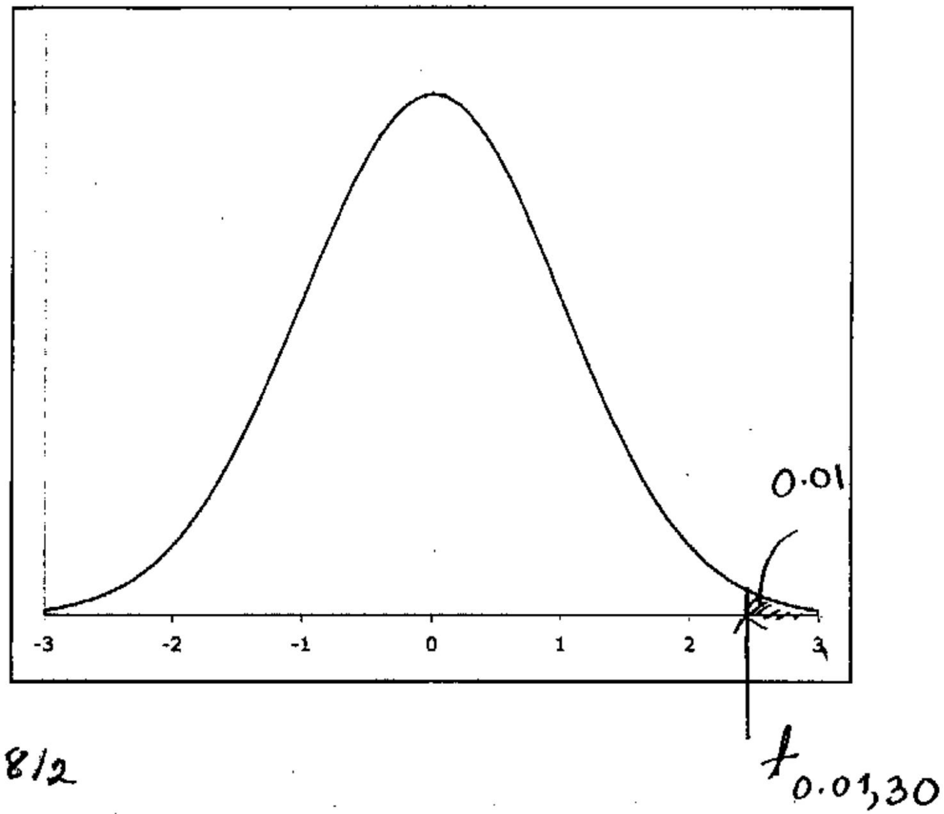
$$\overline{X} \pm t_{\alpha/2;n-1} s/\sqrt{n}$$

Studentin t-jakauman taulukkoarvot $t_{\alpha,df}$ ja $t_{\alpha/2,df}$






0,05
-3
-2
-1
0
1
2
3
$t_{0.05,10} = 1,812$


0.01
-3
-2
-1
0
1
2
3
$t_{0.01,30}$

Esim. 7.6.9. Tiedetään, että eräs kirjailija käyttää tuotannossaan virkkeitä, joiden keskipituus on 32 sanaa. Tutkija lukee erään tekstin, jossa on 30 virkettä. Näiden 30 virkkeen keskipituus on 35,0 sanaa ja keskihajonta 6,8 sanaa. Voisiko teksti olla peräisin kyseisen kirjailijan tuotannosta?

Muodostetaan odotusarvon 95 %:n luottamusväli. Nyt $t_{0,05/2;30-1} = 2,045$ ja luottamusväli $35,0 \pm 2,045 \cdot 6,8/\sqrt{30}$. Saadaan väliksi 32,5 – 37,5, jolle 32 ei kuulu. Päätellään, että teksti ei ole kyseisen kirjailijan tuotantoa.

Esim. Eräs tehdas valmistaa tiiliä, joiden keskipaino pitkän aikavälin seurannassa on ollut 2,000 kg. Erään päivän tuotannosta valittiin satunnaisesti 16 tiiltä. Näiden keskipainoksi saatiin 1,972 kg ja keskihajonnaksi 0,054 kg. Onko keskipainossa tapahtunut muutosta?

Muodostetaan odotusarvon 95 %:n luottamusväli.

Nyt $t_{0,05/2;16-1} = t_{0,025;16-1} = 2,131$ ja luottamusväli

$1,972 \pm 2,131 \cdot 0,054/\sqrt{16}$.

Saadaan väliksi $1,972 \pm 0,029$.

Koska 2,000 kuuluu välille, päätellään ettei muutosta.

Esim. 7.6.7. Esimerkissä 7.6.6, syntymäpituuden tarkastelu, otoksessa 65 poikaa

Descriptives

			Statistic	Std. Error
pituus	Mean		50,95	,245
	95% Confidence Interval for Mean	Lower Bound	50,47	
		Upper Bound	51,44	
	Std. Deviation		1,972	

Luottamusväli on laskettu

$$50{,}95 \pm 2 \cdot 1{,}972/\sqrt{65}, \ (t_{0{,}05/2;65-1} \approx 2)$$
$$50{,}95 \pm 2 \cdot 0{,}245$$

Esim. Lepakoiden tunnistusmatkat, ks.
http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/esimerkit_kaavoihin.pdf

Esim. 7.6.8. Esimerkin 5.1.30 kovuusindeksien erotukset -5, 1, -2, -5, 2,-7, -1, -7, 1, 0, joista keskiarvo -2,3 ja keskihajonta 3,4.

Odotusarvon 95 %:n luottamusväli

$-2{,}3 \pm 2{,}262 \cdot 3{,}4/\sqrt{10}$

$-2{,}3 \pm 2{,}4$

Lisäaineilla ei eroja, koska nolla kuuluu luottamusvälille.

Descriptives

			Statistic	Std. Error
Erotus	Mean		-2,3000	1,08577
	95% Confidence Interval for Mean	Lower Bound	-4,7562	
		Upper Bound	,1562	
	Std. Deviation		3,43350	

7.6.3 Kahden populaation odotusarvon erotuksen luottamusväli

Jos halutaan arvioida kahden populaation odotusarvojen yhtäsuuruutta, niin voidaan arvioida odotusarvojen erotusta $\mu_1 - \mu_2$.

Esim. 7.6.10. Arvio lähiö- ja keskusta-asuntojen keskineliöhintojen erotukselle on (-989,844, -798,862), muodostettu luottamusväli odotusarvojen erotukselle. Koska nolla ei kuulu luottamusvälille, niin päätellään: odotusarvojen olevan eri suuret (keskihinnat eivät samoja).

http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=78

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Neliöhinta	Equal variances assumed	1,235	,268	-18,455	227	,000	-894,35342	48,46101	-989,844	-798,862

Aineisto Asunnot_2006

7.7 Hypoteesien testausta

Tutkimusongelmia:

- Puolueen kannatus?

  Väite: $\pi = 18\ \%$

- Virheellisten komponenttien osuus tuotannossa?

  Väite: $\pi = 4\ \%$

- Kynttilöiden keskimääräinen palamisaika?

  Väite: $\mu = 20$ h

- Asuntojen keskimääräiset neliöhinnat keskustassa ja lähiössä?
  Väite: $\mu_1 = \mu_2$
- Painon ja pituuden välinen lineaarinen riippuvuus?
  Väite: $\rho = 0$
- Opintosuunnan vaikutus kurssiarviointiin?
  Väite: ei riippuvuutta

Tilastollinen hypoteesi on väite populaatiosta, usein populaation jakauman parametrista.

Esim. H_0: $\pi = \pi_0$

H_0: $\mu = \mu_0$

Hypoteesin testaus on väitteen tutkimista otoksen perusteella.

Käytetään sopivaa otossuuretta (testisuuretta), jonka jakauma tunnetaan, kun H_0 tosi.

Esim.

H_0: $\pi = \pi_0$

$$Z = \frac{p - \pi_0}{\sqrt{\pi_0(100 - \pi_0)/n}} \overset{likimain}{\sim} N(0,1)$$, kun H_0 tosi

H_0: $\mu = \mu_0$

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$, kun H_0 tosi

Otossuureen (testisuureen) arvon perusteella H_0 hyväksytään tai hylätään.

Jos testisuureen arvoa pidetään tavanomaisten arvojen joukkoon kuuluvaksi, niin H_0 hyväksytään. Jos arvoa pidetään harvinaisten arvojen joukkoon kuuluvaksi, niin H_0 hylätään.

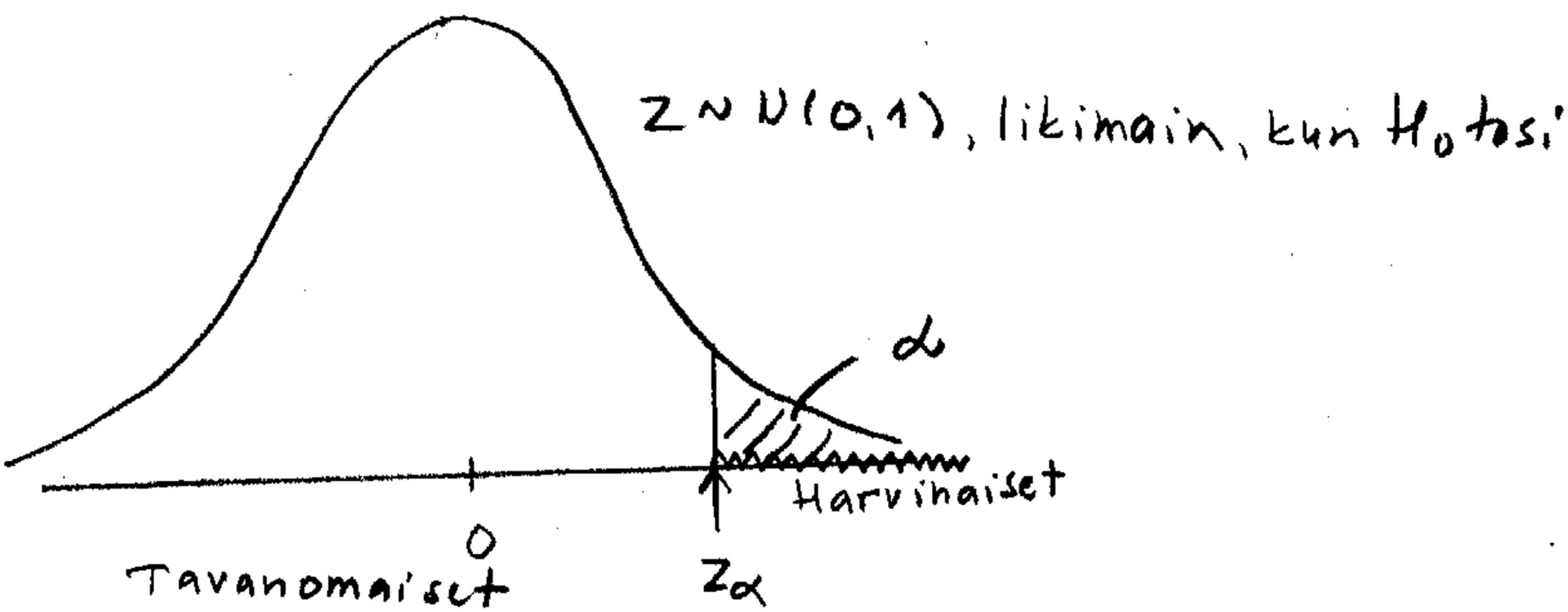
Mikä on harvinaista? Mihin H_0:n hylkääminen johtaa?

Esim. H_0: $\pi = \pi_0$

H_1: $\pi > \pi_0$ vaihtoehtoinen hypoteesi

$$Z = \frac{p - \pi_0}{\sqrt{\pi_0 (100 - \pi_0)/n}} \overset{likimain}{\sim} N(0,1)$$, kun H_0 tosi

Nyt harvinaisiksi arvoiksi katsotaan ”suuret arvot”, suuremmat kuin z_α. Tällöin merkitsevyys- eli riskitaso on α. Usein α = 0,05, 0,025, 0,01 tai 0,001.



Jos H_0 hylätään, niin H_1 hyväksytään.

Miten H_1 asetetaan?

Esim. H_0: $\pi = \pi_0$

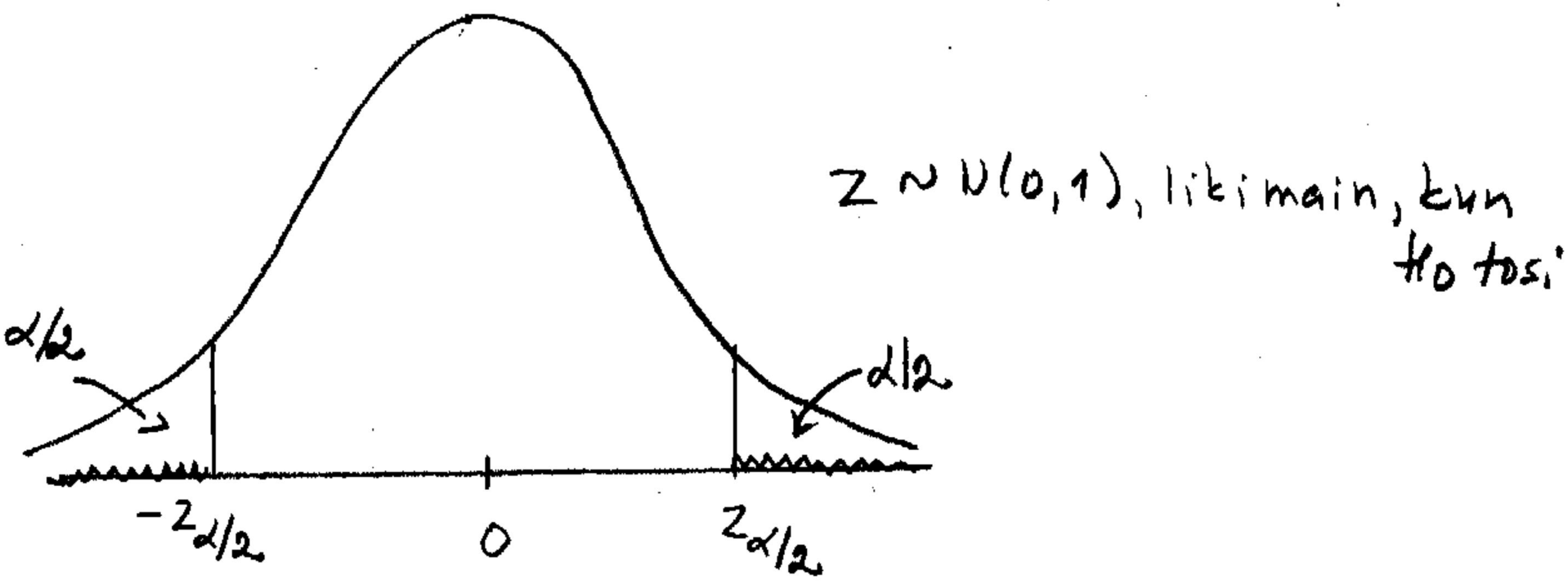
H_1: $\pi > \pi_0$ (yksisuuntainen testi) tai
H_1: $\pi < \pi_0$ (yksisuuntainen testi) tai
H_1: $\pi \neq \pi_0$ (kaksisuuntainen testi)

Kaksisuuntaisessa testissä harvinaisten arvojen joukko on jakauman ”hännillä”.

Sivut 24-28 seuraavalle luennolle

Esim. H_0: $\pi = \pi_0$

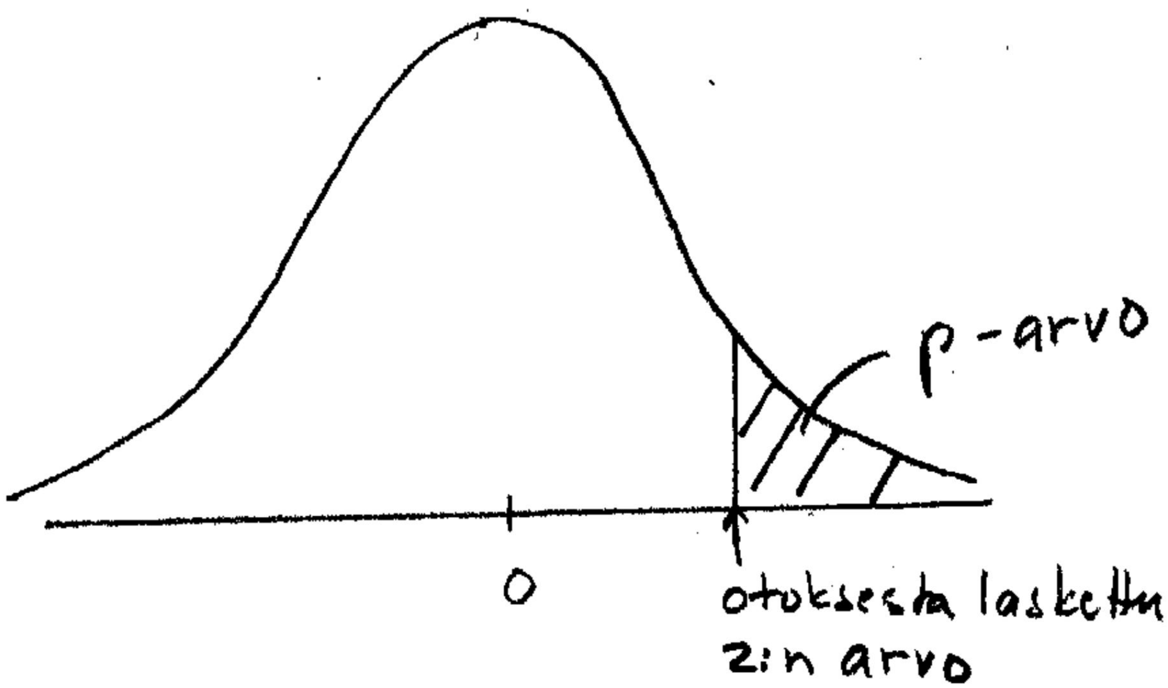
H_1: $\pi \neq \pi_0$ (kaksisuuntainen testi)



Päättely p-arvon perusteella, p-arvo on pienin riskitaso, jolla H_0 voidaan hylätä.

Esim. H_0: $\pi = \pi_0$
H_1: $\pi > \pi_0$

Jos p-arvo on pienempi kuin valittu riskitaso, niin H_0 hylätään.



Testisuureita (seuraavalla luennolla)

H_0: $\pi = \pi_0$

$$Z = \frac{p - \pi_0}{\sqrt{\pi_0 (100 - \pi_0)/n}} \overset{likimain}{\sim} N(0,1)$$, kun H_0 tosi

Esim. 7.7.1. Tarkastellaan erään liikkeen asiakkaita. Halutaan tutkia, ovatko asiakkaista yli puolet naisia. Tehdään 200 asiakkaan satunnaisotos jossa naisia on 113.
Nyt

H_0: π = 50 % ja

H_1: π > 50 %.

Otoksessa naisia 56,5 %. Aineiston perusteella testisuureen arvoksi saadaan

$$z = \frac{56,5 - 50}{\sqrt{50(100-50)/200}} = 1,838 > z_{0,05} = 1,6449 .$$

H_0 hylätään 5 %:n riskitasolla, mutta ei 2,5 %:n riskillä, joten $0{,}025 < \text{p-arvo} < 0{,}05$.

Ks.
http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=81

MTTTP1, luento 9.10.2018

KERTAUSTA TESTAUKSESTA, p-ARVO

- Asetetaan
  H_0
  H_1

- Valitaan testisuure, jonka jakauma tunnetaan H_0:n ollessa tosi.

- Lasketaan otoksesta testisuureelle arvo.

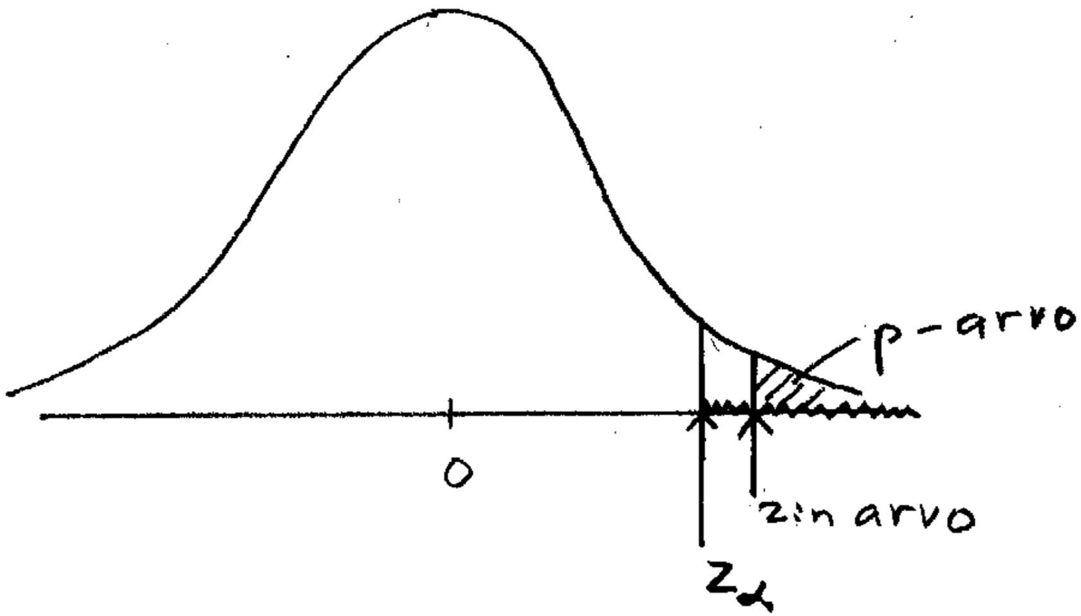
- Kiinnitetään riskitaso α (eli todennäköisyys, että tosi H_0 hylätään). Määritetään testisuureen harvinaisten arvojen joukko, joka riippuu riskitasosta ja vaihtoehtoisesta hypoteesista.

- Hylätään H_0, jos saatu arvo kuuluu harvinaisten arvojen joukkoon, muulloin hyväksytään.

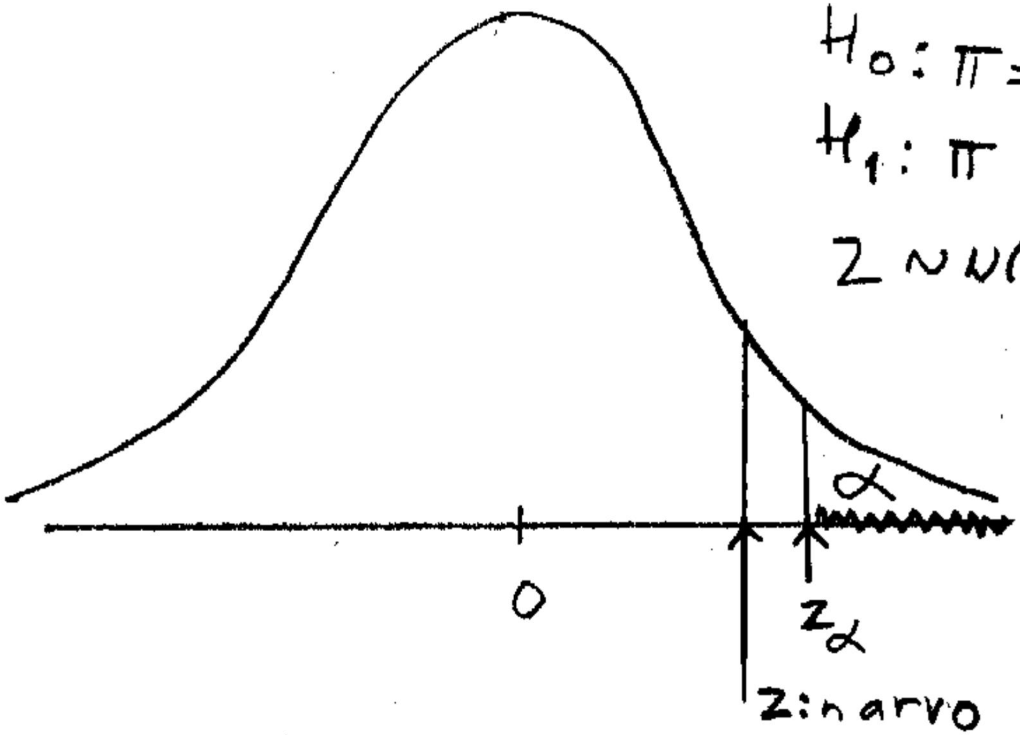
- H_0:n hylkääminen johtaa H_1:n hyväksymiseen.

- Päättely voidaan myös tehdä nk. p-arvon perusteella, p-arvo on pienin riskitaso, jolla H_0 voidaan hylätä. Jos p-arvo on valittua riskitasoa α pienempi, niin H_0 hylätään, muulloin hyväksytään.

H_0: $\pi = \pi_0$, H_1: $\pi > \pi_0$, $Z \sim N(0, 1)$ likimain, kun H_0 tosi.



p-arvo < α,
H_0 hylätään



$H_0: \pi = \pi_0$

$H_1: \pi > \pi_0$

$Z \sim N(0,1)$ likimain, kun H_0 tosi

p-arvo $> \alpha$

H_0 hyväksytään

p-arvo	tulos tilastollisesti
< 0,05	melkein merkitsevä
< 0,01	merkitsevä
< 0,001	erittäin merkitsevä

7.7.1 Prosenttiosuuden testaus (kertausta)

H_0: $\pi = \pi_0$

$$Z = \frac{p - \pi_0}{\sqrt{\pi_0 (100 - \pi_0)/n}} \overset{likimain}{\sim} N(0,1)$$, kun H_0 tosi

Esim. 7.7.1. Tarkastellaan erään liikkeen asiakkaita. Halutaan tutkia, ovatko asiakkaista yli puolet naisia. Tehdään 200 asiakkaan satunnaisotos jossa naisia on 113.

Nyt

H_0: π = 50 % ja

H_1: π > 50 %.

Otoksessa naisia 56,5 %. Aineiston perusteella testisuureen arvoksi saadaan

$$z = \frac{56{,}5 - 50}{\sqrt{50(100-50)/200}} = 1{,}838 > z_{0{,}05} = 1{,}6449$$

H_0 hylätään 5 %:n riskitasolla, mutta ei 2,5 %:n riskillä, joten 0,025 < p-arvo < 0,05.

Ks.
http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=81

Esim. 7.7.2. Eräs puolue väittää, että suomalaisista 40 % kannattaa sen tekemää ehdotusta. Väitteen tutkimiseksi teit kyselyn 1000 henkilölle, joista 360 ilmoitti kannattavansa kyseistä ehdotusta. Onko puolue arvioinut ehdotuksen kannattajat oikein?

Nyt

H_0: π = 40 % ja H_1: π < 40 %

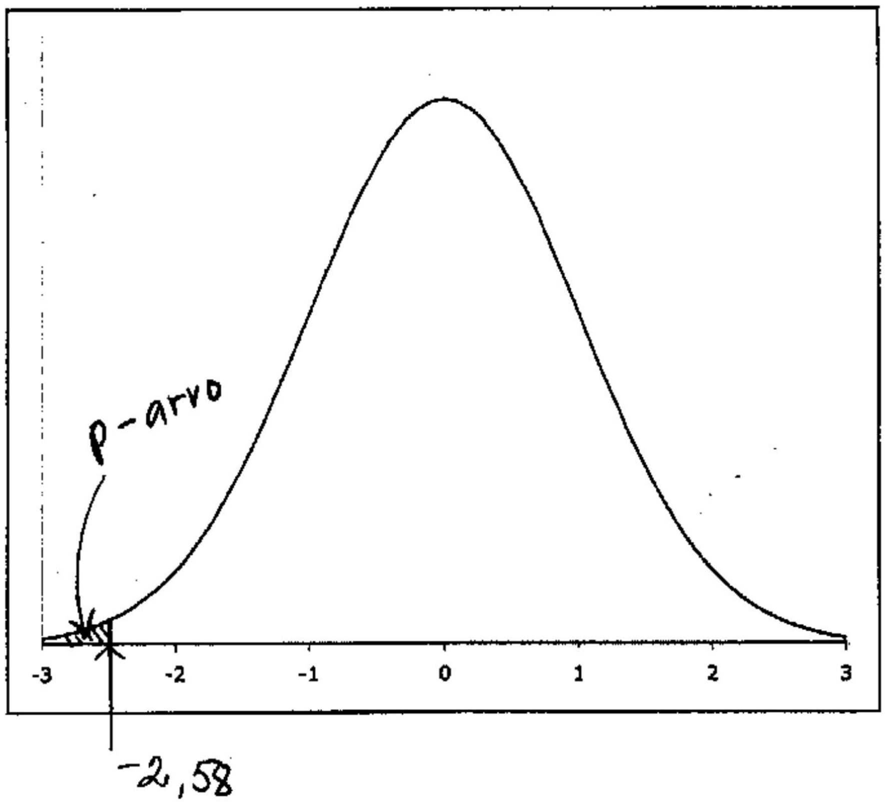
Otoksessa kannattajia 36 %. Testisuureen arvoksi saadaan

$$z = \frac{36-40}{\sqrt{40(100-40)/1000}} = -2{,}58 \quad .$$

Onko tämä harvinaisten arvojen joukkoon kuuluva?

Jos tehdään testaus 1 %:n riskitasolla, niin harvinaisiksi arvoiksi katsotaan lukua $-2{,}3264 = -z_{0,01}$ pienemmät ja täten hylätään nollahypoteesi $(-2{,}58 < -2{,}3264)$.

Pienin riskitaso, jolla H_0 voidaan hylätä, on $P(Z \leq -2{,}58) \approx 0{,}005$, koska $P(Z \leq -2{,}5758) = 0{,}005$.



Vaihtoehtoinen hypoteesi voidaan asettaa myös kaksisuuntaisena, jolloin H_1: $\pi \neq 40$ %.

Tällöin *p*-arvo on $P(Z \leq -2{,}58) + P(Z \geq 2{,}58) \approx 0{,}01$.

Jos valitaan kaksisuuntaisessa testissä tätä suurempi riskitaso (esimerkiksi 0,025), niin nollahypoteesi hylätään ja päätellään, että puolue on arvioinut ehdotuksen kannattajien osuuden väärin.

Esim. Kahvin myyjä väittää, että 15 % kahvin juojista valitsee kahvimerkin hinnan perusteella. Tehdään tutkimus, jossa 250 kahvin juojalta kysytään kahvimerkin valintaa vaikuttavia tekijöitä. Vastanneista 25 valitsi kahvin hinnan perusteella. Uskotko myyjän väitteen?

$H_0: \pi = 15$

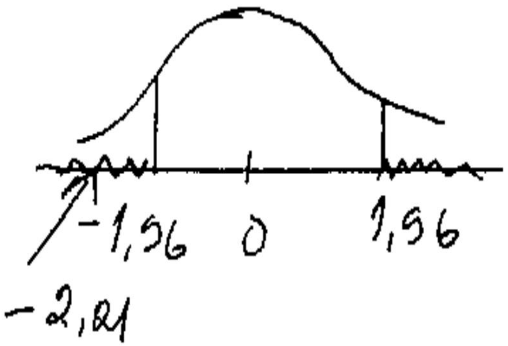
$H_1: \pi \neq 15$

$$Z = \frac{p - \pi_0}{\sqrt{\pi_0(100-\pi_0)/n}} \overset{\text{likimain}}{\sim} N(0,1), \text{ kun } H_0 \text{ tosi.}$$

$$Z_{havaittu} = \frac{10-15}{\sqrt{15(100-15)/258}} = -2,21$$

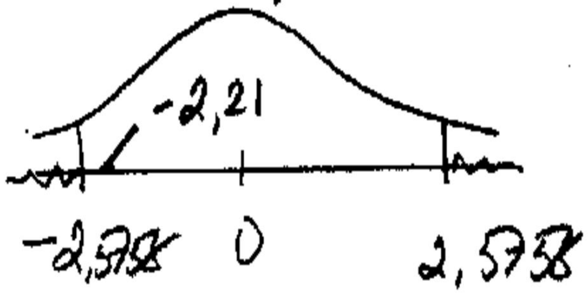
$\alpha = 0,05 \quad \alpha/2 = 0,025$

$z_{0.025} = 1,96$



$\alpha = 0,01, \ \alpha/2 = 0.005$

$z_{0.005} = 2,5758$



H_0 hylätään 5%:n riskitasolla, mutta ei 1%:n riskitasolla.

0,01< p-arvo < 0,05

7.7.2 Odotusarvon testaus

$H_0 : \mu = \mu_0$

$$t = \frac{\overline{X} - \mu_0}{s / \sqrt{n}} \sim t_{n-1}$$, kun H_0 tosi

Esim. 7.7.3. Kauppias väitti, että kananmunien keskipaino on 50 g. Tehdään 36 alkion satunnaisotos ja saadaan $\bar{x}$ = 47, s = 6. Onko kauppiaan väittämään uskomista? Nyt

H_0: μ = 50 g

H_1: μ < 50 g

Saadaan

$$t = \frac{47-50}{6/\sqrt{36}} = -3,$$

joka on pienempi kuin $-t_{0,005;35} \approx -2,75$, joten hylätään nollahypoteesi 0,5 %:n riskitasolla (ei uskota kauppiaan väitettä).

Voidaan arvioida, että p-arvo eli $P(t_{35} < -3) < 0{,}005$.

Vaihtoehtoinen hypoteesi voidaan myös asettaa kaksisuuntaisena eli

H_1: $\mu \neq 50$ g

Jos päättely tehdään 1 %:n riskitasolla, niin harvinaisten arvojen rajoina ovat $\pm t_{0,01/2;35} \approx \pm 2{,}75$, joiden ulkopuolelle -3 jää. Päätellään, että $\mu \neq 50$ g.

Nyt p –arvo = $P(t_{35} < -3) + P(t_{35} > 3) < 0{,}01$.

Esim. 7.7.4. Perunalastujen valmistaja ilmoittaa perunalastupussiensa keskipainoksi 340 g. Tutkitaan väitettä ja tehdään 16 pussin satunnaisotos ja saadaan keskipainoksi 336 g ja keskihajonnaksi 11 g. Voitko uskoa valmistajan väitteen?

H_0: μ = 340 g

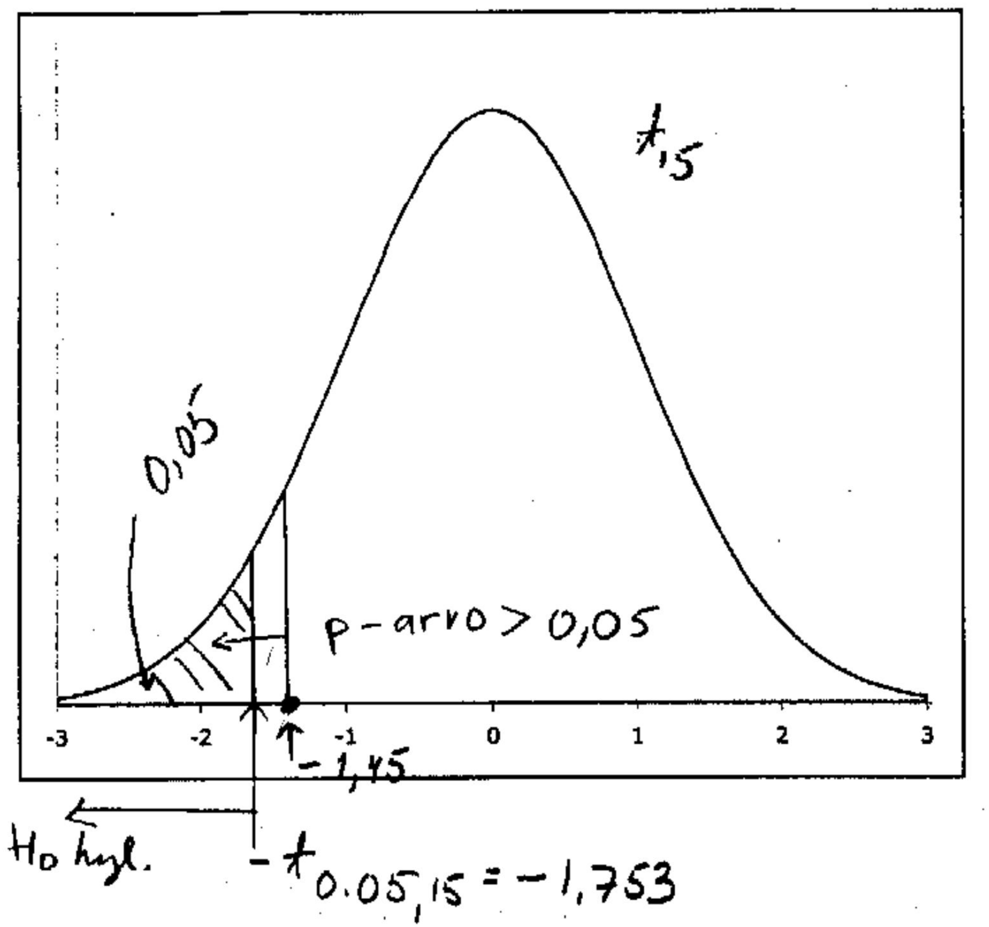
H_1: μ < 340 g

$$t = \frac{336 - 340}{11/\sqrt{16}} = -1{,}45$$

5 %:n riskitasolla harvinaisten arvojen raja on

$-t_{0,05;15}$ = -1,753, joten H_0 hyväksytään.

Nyt *p*-arvo on P(t_{15} < –1,45) > 0,05, graafisesti:



Sama johtopäätelmä tehdään, jos vaihtoehtoinen hypoteesi asetetaan kaksisuuntaiseksi eli H_1: $\mu \neq 340$ g.

Koska $t_{0,05/2;15} = 2,131$, niin H_0 hyväksytään 5 %:n riskitasolla.

Nyt p-arvo = $P(t_{15} < -1,45) + P(t_{15} > 1,45) > 0,10$.

Esim. Lepakoiden tunnistusmatkat, ks. http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/esimerkit_kaavoihin.pdf

7.7.3 χ^2–riippumattomuustesti

Kahden muuttujan välinen riippuvuustarkastelu ristiintaulukon perusteella, asetetaan

H_0: ei riippuvuutta
H_1: on riippuvuutta

H_0 hylätään, jos p-arvo pieni (esim. < 0,05).

7.7.4 Odotusarvojen yhtäsuuruuden testaaminen t-testillä

Tutkitaan kahden populaation odotusarvojen yhtäsuuruutta riippumattomien otosten t-testillä, asetetaan

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

H_0 hylätään, jos p-arvo pieni (esim. < 0,05).

MTTTP1, luento 11.10.2018

7.7.3 χ^2–riippumattomuustesti

Kahden muuttujan välinen riippuvuustarkastelu ristiintaulukon perusteella, kaava (12), hypoteesit

H_0: ei riippuvuutta
H_1: on riippuvuutta

Testisuure χ^2 –riippumattomuustestisuure, joka noudattaa nk. χ^2 –jakaumaa vapausastein (I-1)(J-1). I ja J ristiintaulukon muuttujien luokkien lukumäärät. Ks. χ^2 –jakauman tiheysfunktio https://fi.wikipedia.org/wiki/%CE%A7%C2%B2-jakauma

Esim. 7.7.8

	Taustamusiikki			
Ostettu viini	Ei	Ranskalainen	Italialainen	Yhteensä
Ranskalainen	30 (36%)	39 (52%)	30 (36%)	99
Italialainen	11 (13%)	1 (1%)	19 (23%)	31
Muu	43 (51%)	35 (47%)	35 (42%)	113
Yhteensä	84	75	84	243

Kun ristiintaulukosta lasketaan χ^2-riippumattomuustestisuure, sen arvoksi saadaan 18,279 ja *p*-arvoksi 0,001. Tehdään päättely 1 %:n riskitasolla. Koska *p*-arvo on pienempi kuin valittu riskitaso, niin päätellään taustamusiikin ja viinin valinnan välillä olevan riippuvuutta.

Esim. Tre_myydyt_kolmiot_2010, http://www.sis.uta.fi/tilasto/tiltp_aineistoja/Tre_myydyt_kolmiot_2010.sav

Kolmion kunto * Sijainti Crosstabulation

			Sijainti					
			Keskusta	Kaleva, Amuri, Pyynikki	Hatanpää, Nekala, Epilä, Kissanmaa	Lentävänniemi	Hervanta	Total
Kolmion kunto	Hyvä	Count	23	21	25	27	24	120
		% within Sijainti	76,7%	56,8%	75,8%	71,1%	66,7%	69,0%
	Tyydyttävä tai huono	Count	7	16	8	11	12	54
		% within Sijainti	23,3%	43,2%	24,2%	28,9%	33,3%	31,0%
Total		Count	30	37	33	38	36	174
		% within Sijainti	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4,286[a]	4	,369
Likelihood Ratio	4,220	4	,377
Linear-by-Linear Association	,013	1	,910
N of Valid Cases	174		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 9,31.

Hyväksytään H_0: ei riippuvuutta, koska *p*-arvo (Pearson Chi-Square) on 0,369 > 0,05.

Testiä voidaan käyttää

a) $(I-1)(J-1) = 1$
- $n > 40$
- $20 \leq n \leq 40$, kaikkien teoreettisten frekvenssien (expected count) oltava ≥ 5

b) $(I-1)(J-1) > 1$
- kaikkien teoreettisten frekvenssien oltava > 1 ja enintään 20 % saa olla alle 5

Esim. 7.7.6

Opintojakson työläys * opsuunta Crosstabulation

			opsuunta		
			hallinto	taloust	Total
Opintojakson työläys	työläs	Count	13	16	29
		Expected Count	8,5	20,5	29,0
		% within opsuunta	68,4%	34,8%	44,6%
	sopiva	Count	5	15	20
		Expected Count	5,8	14,2	20,0
		% within opsuunta	26,3%	32,6%	30,8%
	vähätöinen	Count	1	15	16
		Expected Count	4,7	11,3	16,0
		% within opsuunta	5,3%	32,6%	24,6%
Total		Count	19	46	65
		Expected Count	19,0	46,0	65,0
		% within opsuunta	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7,668[a]	2	,022
Likelihood Ratio	8,680	2	,013
Linear-by-Linear Association	7,548	1	,006
N of Valid Cases	65		

a. 1 cells (16,7%) have expected count less than 5. The minimum expected count is 4,68.

Testin käyttöön liittyvät oletukset tällä luokituksella kunnossa, vain 16,7 % odotetusta frekvensseistä alle 5 ja kaikki > 1.

Pienin riskitaso, jolla H_0 voidaan hylätä, on 0,022. Tätä suuremmilla riskeillä H_0 hylätään, pienemmillä hyväksytään.

7.7.4 Odotusarvojen yhtäsuuruuden testaaminen t-testillä

Tutkitaan kahden populaation odotusarvojen yhtäsuuruutta riippumattomien otosten t-testillä, kaava (13), hypoteesit

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$ (tai yksisuuntaisena)

Oletukset t-testin käyttöön liittyen: riippumattomat otoksen normaalijakaumista, joiden varianssit tuntemattomia, mutta yhtä suuret (yhtäsuuruutta voidaan testata).

Esim. 7.7.9

$$H_0 : \mu_{pojat} = \mu_{tytöt}$$
$$H_1 : \mu_{pojat} \neq \mu_{tytöt}$$

Group Statistics

	sex	N	Mean	Std. Deviation	Std. Error Mean
paino	poika	65	3640,46	438,244	54,357
	tyttö	55	3451,27	523,280	70,559

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
paino	Equal variances assumed	,293	**,589**	**2,156**	118	**,033**	189,189	87,765	**15,390**	**362,987**
	Equal variances not assumed			2,124	105,703	,036	189,189	89,069	12,595	365,783

H_0: Populaatioiden varianssit samoja

Testisuure (Levene's Test for Equality of Variances)

Koska *p*-arvo on 0,589 > 0,05, H_0 hyväksytään, t-testin tulokset 1. riviltä.

$H_0 : \mu_{pojat} = \mu_{tytöt}$

Testisuureen arvo on 2,156, kaksisuuntaisen testin *p*-arvo 0,033.

Jos riskitasoksi valitaan 5 %, niin nollahypoteesi hylätään (koska *p*-arvo < 0,05) ja tehdään päätelmä, että tytöt ja pojat ovat syntyessään keskimäärin eri painoisia.

Jos otettaisiin riski, joka olisi pienempi kuin 3,3 % (vaikkapa 1 %) niin tehtäisiin päinvastainen päätelmä.

Yksisuuntaiseen testiin liittyvä p-arvo on 0,033/2 = 0,0165.

Esim. 7.7.11 Tutkitaan, poikkeavatko Hervannan ja Tesoman keskimääräiset neliöhinnat toisistaan, aineisto Tre_myydyt_asunnot_2012

Group Statistics

	Sijainti	N	Mean	Std. Deviation	Std. Error Mean
Neliöhinta	Hervanta	127	1752,6063	456,78817	40,53340
	Tesoma	54	1593,3333	484,13026	65,88178

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Neliöhinta	Equal variances assumed	,009	,926	2,108	179	,036	159,27297	75,55123	10,18732	308,35861
	Equal variances not assumed			2,059	94,993	,042	159,27297	77,35222	5,70924	312,83669

Otoskeskiarvojen ero 159 euroa, t-testin arvo 2,108, johon liittyvä *p*-arvo on 0,036 (kaksisuuntainen testi). 5 %:n riskitasolla päätellään, että keskihinnat eroavat, 1 %:n riskitasolla päätellään, että ei eroja.

Odotusarvojen erotuksen luottamusvälin perusteella odotusarvojen erotuksen arvellaan olevan välillä 10 € – 308 €. Jos päättely tehdään luottamusvälin eikä testisuureen perusteella, niin päätellään odotusarvojen poikkeavan toisistaan, koska nolla ei kuulu luottamusvälille.

Varianssit voitiin olettaa yhtä suuriksi, koska p-arvo on 0,926 (> 0,05).

Esim. Tre_myydyt_kolmiot_2010, http://www.sis.uta.fi/tilasto/tiltp_aineistoja/Tre_myydyt_kolmiot_2010.sav , keskimääräiset koot alueittain, eroja joidenkin alueiden välillä.

7.7.5 Lineaarinen riippuvuus

Populaation korrelaatiokertoimen testaus

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/luentorunko.pdf#page=93

KERTAUSTA

Kaavat esimerkkeineen

http://www.sis.uta.fi/tilasto/mtttp1/syksy2018/esimerkit_kaavoihin.pdf

SEURAAVAKSI

MTTTP5 Tilastollisen päättelyn perusteet, ks.
https://coursepages.uta.fi/mtttp5/syksy-2018/